1-1-1979

Specification and validation of reading skills hierarchies.

Mary Lyn Bourque
University of Massachusetts Amherst

SPECIFICATION AND VALIDATION OF READING SKILLS HIERARCHIES

A Dissertation Presented

By

MARY LYN BOURQUE

Submitted to the Graduate School of the
University of Massachusetts in partial fulfillment
of the requirements for the degree of

DOCTOR OF EDUCATION

May 1979

School of Education

SPECIFICATION AND VALIDATION OF

READING SKILLS HIERARCHIES

A Dissertation Presented

By

MARY LYN BOURQUE

Approved as to style and content by:

Ronald K. Hambleton,
Chairperson of Committee

Hariharan Swaminathan, Member

J. Michael Royer, Member

Mario Fantini, Dean
School of Education

A C K N O W L E D G E M E N T S

When candidates for the doctorate degree reach the final stages of completing the University's requirements, they must reflect on the years of study completed and remember with fondness and gratitude the significant people whose lives touched theirs in the learning process and without whom there would be no commencement—no beginning.

I am deeply grateful to each of the members of my Committee, Ronald Hambleton, Hariharan Swaminathan and J. Michael Royer, for their inspiration and support during my graduate program. Dr. Hambleton, Chairperson, was particularly instrumental in providing me with not only thorough academic experiences, but also a broad variety of field-based opportunities while in residence at the University. For this, as well as for his encouragement and professional guidance I am personally grateful.

I would also like to extend a special thanks to Leah Hutten. Without her untiring efforts at programming and data analysis this study could not have gone forward.

I am most happy to formally acknowledge the unquestioning moral and financial support of my parents throughout my years of study. Without their sacrifices and personal

efforts on my behalf, this goal would never have become a reality. Likewise, I must remember with fondness my undergraduate experiences at Emmanual College, Boston. I am sure that without the initial inspiration during those formative years, graduate study would not have been one of my aspirations.

To my friends who gracefully sustained my social delinquency during the study, and to those professional colleagues who encouraged me to persevere to the end I am also grateful. I am equally indebted to Delta Kappa Gama International, Alpha Chapter, for their generous financial support of this study.

Finally, a special thank-you must be given to Bernie McDonald who typed the final manuscript.

ABSTRACT

Specification and Validation of
Reading Skills Hierarchies

February 1979

Mary Lyn Bourque, A.B., Emmanuel College
M.Ed., Boston College, Ed.D., University of Massachusetts

Directed by: Professor Ronald K. Hambleton

There is a growing concern among practitioners and academicians alike for the numbers of elementary and secondary school pupils who are unable to read and compute at the termination of their school experiences. This concern has sparked a growing interest in instructional and measurement systems research. One topic which is of particular interest is that of learning hierarchies in reading and language arts. Learning hierarchies have been the structure of basal programs as well as reading management systems for a number of years. However, one of the weak aspects of both is the lack of research which focuses on the specification and validation of such hierarchies.

The purpose of this study is to compare two empirical methodologies for establishing hierarchical relationships, viz., the Dayton and Macready model (1976), and the White and Clark procedure (1973), with an a priori hierarchy established by content area specialists.

Eight phonics skills and eight structural analysis skills were selected from the test battery, *The Reading Skills Inventory: A Criterion-Referenced Assessment* (Hambleton, 1975).

In order to establish an a priori hierarchy based on the judgment of experts in the field of reading a sample of 23 content specialists was asked to respond to a pair-wise comparison task. Each respondent examined 56 pairs of objectives resulting from two 8-objective clusters: one cluster of phonics skills, and a second of structural analysis skills. The resulting hierarchies were then compared to those produced using empirical data based on the administration of four criterion-referenced test levels to approximately 14,000 elementary school children in an urban setting.

Initially the Dayton and Macready model for specifying a hierarchy utilizing a maximum likelihood solution was applied. Secondly, the White and Clark procedure, a pair-wise comparison method having a "test of inclusion" significance test, was applied. This procedure can accommodate multi-item data-sets for each objective in the hierarchy, and, as a result, can estimate the probability of a randomly selected examinee having answered zero, or one or more items correctly for any objective-pair.

Hierarchy specification by content experts revealed an overwhelming lack of agreement among reading specialists on the hierarchical relationships among reading skills and objectives. In part, this is due to the lack of clarity and preciseness of articulating reading objectives in behavioral terms.

Specification of the hierarchies based on empirical data via either probabilistic model produced more stable results. Several critical problems were identified in using both models, and solutions are proposed.

The results of the study indicate that it is quite possible to specify and validate hierarchical relationships. First, validated hierarchical patterns among reading skills should result in improved instructional sequences. If prior acquisition of certain skills is necessary to the posterior acquisition of alternate skills then proper curriculum sequencing becomes critical in the instructional design process. Secondly, validated hierarchies should allow for more efficient and effective diagnosis and prescription. This is particularly important when the practitioner in reading is faced with the problems of remediation. Finally, hierarchical relationships among instructional objectives should foster the development of tailored testing programs as well as improve more formative pupil evaluation procedures.

TABLE OF CONTENTS

Page

APPENDICES

LIST OF TABLES

Table

Page

LIST OF FIGURES

CHAPTER I

INTRODUCTION

1.1 Background

The most recent headlines in the national news reveal that College Board SAT scores are down again this year. Verbal score declined 14 points confirming the adage that "Johnny still can't read," and suggesting that perhaps Johnny is becoming more illiterate with each succeeding generation. Blue ribbon panels are commissioned with great regularity to investigate the illiteracy problem, only to find that the results of such research merely diagnose in greater detail the symptoms of a "national disease." And so the great debate over why all this is true and how schools can and should make a difference goes on.

In an effort to solve the illiteracy problem psychologists have diagnosed, analyzed, synthesized and evaluated the learning process, while specialists have tried to apply this theory to the process called reading.

Each decade sees a new instructional panacea. Competing methodologies have emerged, out of which systems of teacher-training and staff development have evolved. A cursory review of the IRA Annual Convention Program

would reveal great diversity of both philosophy and technique in the teaching of reading. However, despite such diversity, there appears to be one common practice that has survived, and indeed, been perpetuated throughout the debate: that the reading process is a series of sequential and ordered learning tasks forming a hierarchy, and that within the reading hierarchy there is a positive transfer of learning from lower level tasks to higher level tasks.

Publishers of reading text books have capitalized on this assumption by offering the consumer reading series that are designed to unfold the reading hierarchy in a methodical way. These are called basal readers and are used almost universally to pace the learner through a series of hierarchical tasks ranging from readiness skills to decoding and structural analysis skills. Publishers of norm-referenced achievement tests have paralleled the basal text approach. Lower levels of reading achievement tests are designed to measure primarily readiness skills, while a content analysis of upper test levels reveals an emphasis on higher order decoding and comprehension skills. Progression from simple tasks to more complex ones is intuitively appealing and is a reflection of reality in many instances of both formal and informal learning. The works of Piaget and Gessel attest to that fact. However, whether the notion of

developmental learning can be generalized to the field of reading is a serious question that has yet to be answered.

1.2 Purpose of the Study

Because the issue of sequencing reading skills is a critical one having serious implications for both the theoretician and the practitioner, this study has focused on the feasibility of specifying and validating such hierarchies using several alternate methodologies.

Specifically, the purpose of this study was threefold:

1. To establish, using empirical test data, whether or not a hierarchy exists among selected reading skills;

2. To establish the direction of the hierarchy and the strength of the relationship among the component reading skills;

3. To compare several empirical methodologies for establishing hierarchical relationships.

This study was limited to the empirical mode of establishing and validating a reading hierarchy among selected, low-level decoding skills. When a hierarchy was found to exist, the study pursued a more detailed analysis of the hierarchy. That is, an attempt was made to establish the direction and strength of the relationship among the component skills suggested by the data.

Finally, various psychometric methodologies for validating learning hierarchies were compared and the advantages and disadvantages of each approach where possible were pointed out.

Hierarchies are usually established and validated in one of three ways or combinations thereof. First an a priori ordering can be generated. In the case of reading, this type of ordering is usually arrived at deductively based on reading theory. The basal reader and reading achievement tests are probably the best examples of this type of ordering. One of the initial test-development procedures involves a thorough examination of the major reading series for common behavioral content and sequence. Normally this scope and sequence has been arrived at through the collective judgment of professionals in the field of reading. In many instances the ordering is subsequently verified by the reading practitioner.

A second approach to generating reading hierarchies utilizes psychometric data and is classified as empirical. This type of ordering is usually based on test data and is confirmatory in nature. That is, an existing a priori ordering is measured using appropriately generated and sampled item-sets. The data are expected to be confirmatory of the hierarchy.

Finally, a third approach can be classified as experimental. That is, given an existing (or hypothesized) a priori ordering, an instructional unit is designed to teach to the hierarchy. Subjects are then tested on a pre/post instructional basis for skill acquisition in the hierarchy; hierarchical relationships are then established.

1.3 Educational Significance of the Study

The dynamics of the reading process is a very complex and yet unsolved mystery of the human mind. During this century there have been landmark contributions to a vast field of literature and theory in the works of Huey (1908), Neisser (1967) and others. Wolf (1977) has pointed out that only recently, however, has the concept and definition of reading been expanded beyond the conventional understanding of decoding the written word. Reading research is now redirecting its efforts and focusing on the psychological processes involved: memory storage, systems of attention and bilateral perception. This in turn has triggered a closer scrutiny of reading instruction and student performance. The criterion-referenced assessment movement has also had its impact on reading research. The design of reading instruction has become objective-based as has the accompanying testing programs. However, much of the research in objective-based programs has depended on hierarchies in areas other than the field of

reading. This is in part due to the fact that other disciplines such as mathematics and science more easily lend themselves to decomposition into discrete learning units called objectives. Therefore, one of the primary contributions of this study was to explore the generalizability of learning hierarchy theory to the area of reading and the applicability of several psychometric techniques for validating such a hierarchy. It is hoped that the results of the study will contribute to the field of instructional design insofar as validated hierarchical relationships can be useful in packaging instructional units in reading. Concomitantly, criterion-referenced reading testing programs could be significantly streamlined if empirical evidence was available on the hierarchical structure of objective-based reading programs.

CHAPTER II

A REVIEW OF THE LITERATURE

2.1 Theoretical Considerations in Criterion-Referenced Measurement

This literature review will be developed in two parts. Section 2.1 will deal with some general aspects, theoretical considerations, in criterion-referenced measurement. Because this section is merely background against which to set the real issue, viz., hierarchy validation, no attempt is made to present a complete and thorough exposition of the topics included. Section 2.2, however, will present an in-depth review of the methodological issues relevant to learning hierarchy specification and validation.

2.1.1 Types of Hierarchies

When examining learning hierarchies it is possible to identify four primary types of ordering among the components: (a) completely ordered, (b) completely independent or unordered, (c) weakly ordered, and (d) mixed ordering (combinations of a, b and c). Each of these four primary types can be either strictly linear or branching. Gagné (1970) defines a learning hierarchy as ". . . an entire

set of capabilities having an ordered relation to each other (in the sense that in each case prerequisite capabilities are presented as subordinate in position, indicating that they need to be previously learned). . ." (p. 238). Baker and Hubert (1977) have defined completely ordered components of a hierarchy as equivalent: that is, O_i must be passed in order to pass O_j and conversely.

A learning hierarchy which has completely unordered or independent elements assumes no interrelationship among them. In other words, mastery of one skill is totally independent of mastery of the other skill. Such a relationship might exist, let us say, between two mathematics skills such as solving a quadratic equation and finding the area of a two-dimensional figure.

Weak ordering among elements in a hierarchy is defined as prerequisite ordering by Baker and Hubert (1977). That is, such a relationship among any given pair of elements facilitates learning of the higher skill if mastery of lower skill obtains. Gagné (1970) discusses this type of ordering in terms of facilitating the transfer of learning from the lower-order skill to the higher-order skill.

A mixed system involves a combination of one or more of the primary types. Figure 1 provides the reader with a graphical display of these four types of ordering.

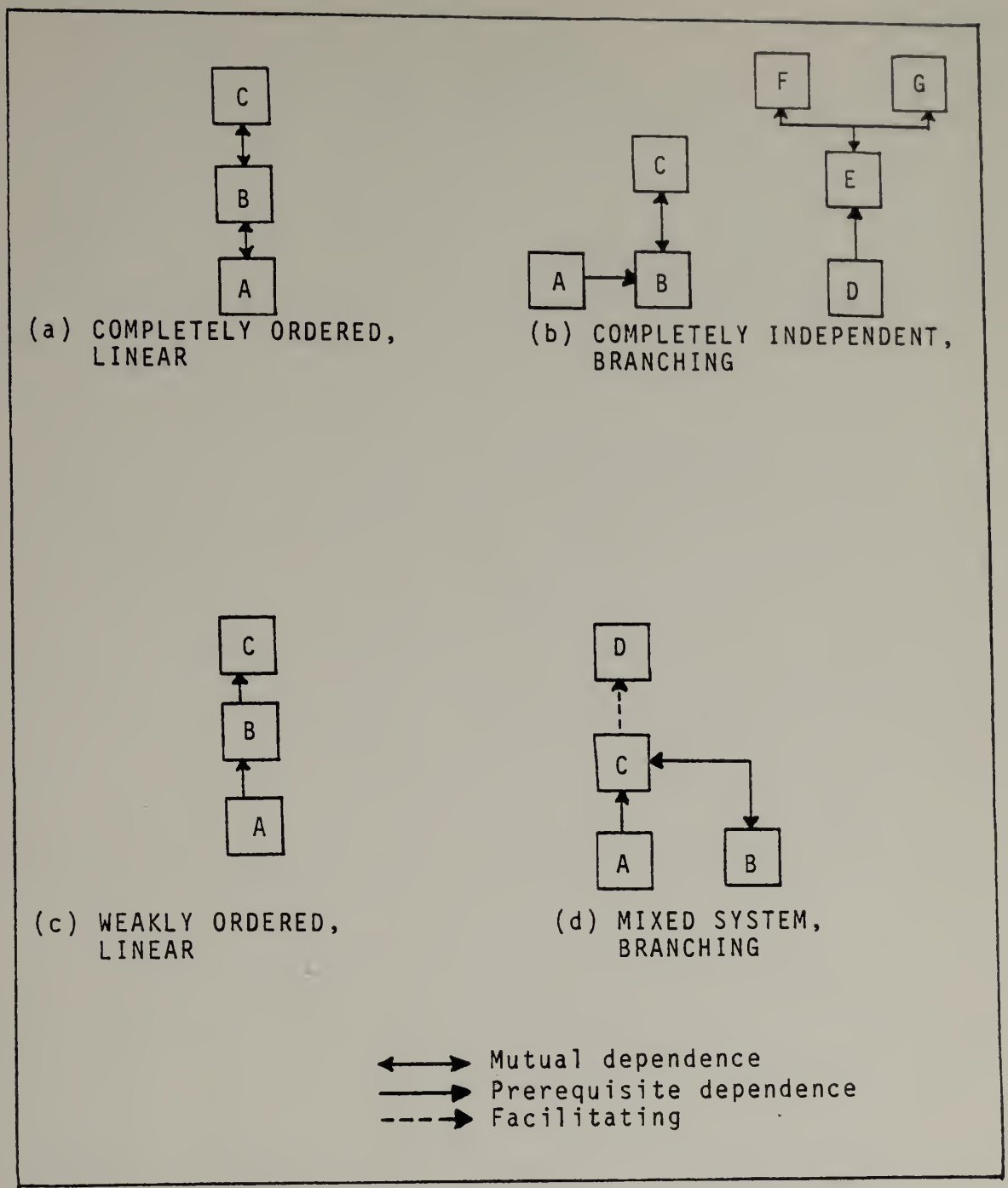

Figure 1. Examples of the four primary types of both linear and branching hierarchies.

2.1.2 Reliability of Test Item Scores

One of the major problems in hierarchy specification and validation is attaining precision in measuring whether or not an examinee has mastered the objectives in the hierarchy. All of the validation methodologies are highly contingent upon reliable assessments of mastery. Therefore, the researcher must ask the question, "If the measurement is not perfectly reliable, how many items are necessary in order to approach 'good' reliability"? First, let us dispense with the ideal situation of "perfectly" reliable items. If this is the case, then one item per objective is both the necessary and sufficient condition for judging mastery on each objective. Normally, however, the ideal does not prevail. This precipitates a problem immediately, since most of the hierarchy validation procedures are designed to use one item per objective to assess mastery. It is possible, however, to use a multi-item average score, or to make a systematic or random selection of items for each objective.

2.1.3 Cutting Scores

Assuming that items are not perfectly reliable, and also assuming that multi-item measures of each objective in the hierarchy are more desirable than single-item measures, the question is what should the cutting scores be which determine the master/non-master decision?

Setting standards is a complex process and there are any number of ways in which a solution can be generated. However, all of the methods for determining cutting scores have one aspect in common: they all involve human judgment. In other words, although the procedures may be quite technical or sophisticated, in the final analysis it becomes an arbitrary decision about what constitutes mastery and what does not.

The preceding section constitutes merely general background against which the primary focus of this review is set. The remaining section will provide the reader with a detailed analysis of the technology of hierarchy specification and validation.

2.2 Methodological Considerations in Validating Learning Hierarchies

There have been many significant contributions to the literature on hierarchy specification and validation over the past two decades. This discussion will review the major published studies, focusing on the techniques employed, the inherent weaknesses in the methodologies if any, and the primary results which catalyzed further research. Because the similarities and differences among the methods can be subtle, Table 1 displays an abridged review of each method discussed, with the hope that the reader may find it helpful.

Table 1

An Abridged Review of Hierarchy Validation Methodologies, 1961 - 1977

Author	Statistical Techniques Employed	Weaknesses in the Model
Gagne and Paradise 1961	PROPORTION OF POSITIVE TRANSFER (PPT) $PPT = \frac{(11) + (00)}{(11) + (00) + (01)}$	1. Unable to estimate errors of measurement since there is only one item per learning set. 2. Forgetting parameter not considered. 3. Inflated PPT index leads to invalid connections in the hierarchy. 4. Lacks sampling distribution and thus, generalizability.
Resnick and Wang 1969	MULTIDIMENSIONAL SCALOGRAM ANALYSIS (MSA) $Rep = 1 - \frac{(01)}{(11) + (00) + (01) + (10)}$	1. Tested hierarchy without any prior instruction. 2. Forgetting parameter not considered. 3. Unknown sequence of skill acquisition. 4. Small sample size 5. MSA restricted to independent linear scales. 6. Lacks sampling distribution and thus, generalizability. 7. Unable to estimate errors of measurement.
Airasian 1971	CONDITIONAL ITEM DIFFICULTY INDEX $P_{i+1 \mid i} = \frac{p_{i+1}}{p_i}$ where p_i = proportion who attain expected response pattern	1. Highly dependent on item validity. 2. Sensitive to sequencing of instruction. 3. Unable to estimate errors of measurement. 4. Forgetting/guessing parameters omitted.
Capie and Jones 1971	Phi-correlation coefficient	1. Estimations are highly dependent on test validity and reliability. 2. Existence of a hierarchy must yield high phi values; but high phi values do not necessarily establish a hierarchy.

Table 1 (continued)

Author	Statistical Techniques Employed	Weaknesses in the Model
Walbesser and Eisenberg 1972	DEPENDENCY TEST RATIOS: Consistency Ratio = $\frac{(11)}{(11) + (10)}$ Adequacy Ratio = $\frac{(11)}{(01) + (11)}$ Completeness Ratio = $\frac{(11)}{(11) + (00)}$ Necessity Ratio = $\frac{(00)}{(00) + (10)}$ Inverse Consistency Ratio = $\frac{(00)}{(00) + (01)}$	1. Errors of measurement suppress both C/R and N/R. 2. N/R particularly sensitive to the number of failures on the test. 3. Lacks generalizability. 4. Suggests invalid connections because of high ratios. 5. Indices measure how easy the test was for the examinees.
Bart, Airasian and Krus 1973	ORDERING-THEORETIC MODEL Percent of disconfirmatory response patterns (01) with pre-established tolerance levels.	1. Limited to pairwise analysis. 2. Depends upon pre-established tolerance levels for dealing with error (deterministic rather than probabilistic).
White and Clark 1973	PROBABILISTIC MODEL Estimates the probability that the examinee will be classified in the (01) cell, which is defined as the sum, over the four possible groups, of an examinee's probability of being in each group (P_0, P_I, P_{II}, P_B), multiplied by the conditional probability of members of the group answering the relevant number of items correctly: $P_{On} = P_0(1-\theta_b)^n \theta_d^n + P_I(1-\theta_a)^n\theta_d^n$ $+ P_{II}(1-\theta_b)^n \theta_c^n + P_B(1-\theta_a)^n \theta_c^n$	1. Valid connections may be rejected if sample size is too large. 2. Tests pairwise connections only.
Dayton and Macready 1976	MAXIMUM LIKLIHOOD MODEL A class of probabilistic models employing maximum liklihood estimates: $P_{(u)} = \sum_{j=1}^{q} P(u \mid v_j) \cdot \theta_j$, where $P(u \mid v_j) = \prod_{i=1}^{k} \alpha_i^{a_{ij}}(1-\alpha_i)^{b_{ij}} \beta_i^{c_{ij}} (1-\beta_i)^{d_{ij}}$ where α_i = guessing parameter and β_i = forgetting parameter	1. Limited to one item per skill.

2.2.1 Gagné and Co-workers

Gagné (1962) defined "knowledge" as that "inferred capability which makes possible the successful performance of a *class of tasks* that could not be performed before the learning was undertaken" (p. 355). The order in which the *class of tasks* is unfolded before the learner, or the order in which the *class of tasks* is successfully mastered by the learner becomes a critical issue for both the curriculum and instructional planner as well as the psychometrician. Gagné and his co-workers (Gagné, 1962, 1968; Gagné, Mayor, Garstens & Paradise, 1962; Gagné & Paradise, 1961) addressed the problem of hierarchy validation with the primary question, "What must an individual know or be able to do in order to achieve successful performance on this class of tasks, assuming that he is given instructions indicating the form of the desired response and stimulus definitions?" (p. 4). Gagné's model is based on the theory that the learner cannot successfully perform a task higher in the hierarchy until s/he has mastered each of the prerequisite lower tasks. And so, beginning with the hypothesized highest task in the hierarchy, the primary question is asked, and then reiterated for each successive lower task. This process is continued until one reaches a point where one can be relatively certain of basic skill acquisition by the population. Once the hypothesized hierarchy has been

structured, an instructional unit is prepared to teach the skills. Upon completion of instruction, examinees can be divided into four categories:

1. Those who pass the higher skill and all lower skills (11).
2. Those who fail the higherskill and at least one of the lower skills (00)
3. Those who fail the higher skill and pass at least one of the lower skills (10)
4. Those who pass the higher skill and fail at least one of the lower skills (01)

The categories described are represented graphically in Figure 2.

		HIGHER SKILL	
		Fail	Pass
LOWER SKILL	Pass	10	11
	Fail	00	01

Figure 2. A 2x2 matrix of the status of examinees who pass or fail lower and higher skills in the hierarchy.

Ideally, the 01 category should be the null set. That is, if positive transfer of learning is operating then there should be no examinees who would pass a higher order skill but fail a lower order one. The index "proportion of

positive transfer" is the ratio of cells (11) + (00) to cells (11) + (00) + (01). Quite obviously the ratio approaches unity as cell (01) approaches zero. Also, Gagné estimated that the values of the index at chance level would be between .25 and .50 (Gagné & Paradise, 1961, p. 9). However, White (1973, p. 363) has demonstrated that completely independent skills in a hierarchy whose correlation is zero can achieve an index value of .67, clearly above chance level by Gagné's definition.

White (1973) has also pointed out other weaknesses in the model. First, only one item per skill was used in the study. Consequently, it was not possible to estimate the errors of measurement associated with the exceptions to the hierarchy. Therefore, one is left in a state of indecision relative to establishing whether or not the observed exceptions reflect true deviations in the hierarchy or chance error in the measurement of true connections.

Secondly, the measurement of each skill was delayed until completion of the instructional unit. This allowed for forgetting effects to impact on the results. Pupils could have learned both lower and higher skills and forgotten the lower ones before testing occurred. This would increase the freqeuncy count in the (01) cell thus yielding lower indices which could invalidate true connections in the hierarchy.

Thirdly, Gagné's model can lead to invalid connections in the hypothesized hierarchy. This occurs because the nature of the model is one that is confirmatory. That is, the "proportion positive transfer" index, which may be spuriously high, can only confirm connections which are postulated; it cannot suggest connections which may have been overlooked.

Finally, since Gagné's method lacks a sampling distribution, the generalizability of the results is quite dubious. There is no evidence that would indicate the variability of the index from one sample to another.

2.2.2 Resnick and Wang

Several years later Resnick and Wang (1969) applied the theory of Guttman scalogram analysis to the validation of learning hierarchies. Their approach was fraught with as many insoluble problems as was Gagné's. Guttman scales, devised for quite dissimilar purposes, enjoy one unique property: the "parallelogram pattern" of responses, which is both the necessary and sufficient condition for a perfect scale. "The existence of such a 'perfect' scale, or an acceptable approximation to it, is taken to confirm the existence of a behavior hierarchy" (Resnick & Wang, 1969, p. 17). The statistical test employed was the conventional coefficient of reproducibility,

$$\text{Rep} = 1 - \frac{(01)}{(11) + (00) + (01) + (10)}$$

where the fraction is simply the ratio of the number of errors to the total number of responses. Conceptually, the coefficient of reproducibility is really a measure of the degree to which the scale in question approaches the perfect Guttman scale. The Guttman coefficient used here is quite different from Gagné's statistic insofar as the coefficient of reproducibility applies to the hierarchy as a whole and not to individual connections as does Gagné's index. Resnick and Wang used the Lingoes (1963) modification of scalogram analysis which allows for the generation of multiple scales and thus, branching hierarchies.

However, while in some respects the Resnick and Wang approach may represent an improvement in the technology of hierarchy validation, their research has been criticized by more recent contributors to the field (cf. Airasian & Bart, 1975; White, 1973). First, the Resnick and Wang study tested kindergarten students on skill acquisition without having provided any prior instructional exposure. With this approach the test is merely a measure of preschool achievement, and in no way does it validate or reject a given hierarchy. Consequently, the sequence of skill acquisition is unknown and the possibility of random forgetting is largely ignored. Additionally, a very small sample size was used. The general lack, therefore, of these experimental controls leads only to a state of

uncertainty about the hierarchy. Moreover, although MSA can generate multiple scales, they are restricted to being independent and linear. That is, once the analysis has identified a skill as "scaleable" with respect to other higher and/or lower skills, that selected skill may not appear in any other place. This, of course, places severe limitations on the technique and its ability to handle branching within a hierarchy. To overcome this limitation it becomes necessary to test every possible linear scale separately—a cumbersome and costly process.

Finally, as in the Gagné model, there is no sampling distribution because the measurement is limited to one item per skill. Therefore, there is no way of estimating errors of measurement and of generalizing beyond the sample.

2.2.3 Airasian

The model suggested by Airasian (1971) is based on the frequency with which examinees attain a true score (expected) pattern of responses. That is, in a 4-item test with 16 total possible response patterns, only n+1 patterns can be assigned a probability of occurrence greater than zero. Those patterns are 0000, 1000, 1100, 1110 and 1111. For each of the four items, therefore, a ratio of the number of examinees who attain an expected pattern with that item correct to the total number of examinees who attain true score (expected) patterns is

calculated. So, for item 1, the ratio is

$$P_1 = \frac{P_{1000} + P_{1100} + P_{1110} + P_{1111}}{P_{0000} + P_{1000} + P_{1100} + P_{1110} + P_{1111}}$$

Similarly, for item 2, the conditional item difficulty index is

$$P_2 = \frac{P_{1100} + P_{1110} + P_{1111}}{P_{1000} + P_{1100} + P_{1110} + P_{1111}}$$

The conditional item difficulty index can be stated generally as

$$P_{i+1 \mid i} = \frac{P_{i+1}}{P_i}$$

where P_i = proportion of examinees who attain the expected or true score response patterns

Airasian admits to at least two serious confounding influences on his model: (a) item validity is critical, and (b) the index is highly sensitive to the sequencing of instruction. Item validity is a problem with each of the models discussed thus far. In addition, the Airasian model does not solve the problem discussed earlier relative to estimating errors of measurement; it does not provide for a sampling distribution; and it largely ignores the forgetting and/or guessing parameters.

2.2.4 Capie and Jones

The work of Capie and Jones (1972) was primarily a comparative study of the dependency test ratios (cf. section 2.2.5) with the phi-correlation coefficient. They based their argument on the fact that the skills in a true hierarchy should clearly demonstrate increasing correlational patterns if one were to compute product moment correlations between individual skills and the total number of skills mastered. This concept is at least intuitively appealing. They also suggest that a parallel set of correlations be derived for a transfer test, based on the premise that increased mastery of skills in the hierarchy must necessarily result in increased transfer of learning. Consequently, increasing correlational patterns should be in evidence for both the transfer test as well as the individual skills if it is a true hierarchy. To compensate for the unreliability of the tests, phi can be corrected for attenuation.

Capie and Jones, by their own admission, are critical of their model. First, the technique is highly contingent upon the validity and reliability of the test. If test validity is questionable, or if the tests are so short that reliability is equally dubious, then the technique is severely weakened. More importantly, however, the testing sequence is based on the arbitrary judgment of the researcher. To the extent that such judgment is on target,

then the testing sequence will reflect all the possible prerequisite skills. This is not always the case, however.

White (1974c) is quite critical of the Capie and Jones' procedure.

> Capie and Jones advocate the establishment of a hierarchy by calculating phi-correlation coefficients for each pair of skills, and, where the coefficients are significantly different from zero, placing the skills in order of difficulty. Although the existence of a hierarchical relation between two skills implies a positive correlation and a difference in difficulty between them, the reverse is not necessarily true. Capie and Jones' criteria are necessary but not sufficient conditions for a valid hierarchy. Use of their criteria alone can lead to a hierarchy which contains superfluous skills, and superfluous connections between skills. (p. 64)

2.2.5 Walbesser and Eisenberg

The dependency test ratios of Walbesser and Eisenberg (1972) are not unlike Gagné's proportion of positive transfer. Walbesser proposes five ratios for specifying hierarchical relationships:

1. Consistency ratio: ratio of examinees mastering the higher skill to those who have mastered the lower skill;
2. Adequacy ratio: ratio of examinees mastering the lower skill to those who have mastered the higher skill;

3. Necessity ratio: ratio of examinees who failed the lower skill and also failed the higher skill;
4. Completeness ratio: ratio of examinees who mastered the higher skill to those who both mastered or failed both skills completely;
5. Inverse Consistency ratio: ratio of examinees who failed the lower skill to those who mastered the higher skill.

These relationships can be represented as,

$$\text{Consistency Ratio} = \frac{(11)}{(11) + (10)}$$

$$\text{Adequacy Ratio} = \frac{(11)}{(01) + (11)}$$

$$\text{Completeness Ratio} = \frac{(11)}{(11) + (00)}$$

$$\text{Necessity Ratio} = \frac{(00)}{(00) + (10)}$$

$$\text{Inverse Consistency Ratio} = \frac{(00)}{(00) + (01)}$$

Walbesser has arbitrarily set a lower limit of .85 for the consistency, adequacy and completeness ratios. If these ratios do not exceed this threshold value then a valid hierarchical connection cannot exist.

Both Capie and Jones (1971) as well as White (1974c) have been critical of the Walbesser model. First, Capie and Jones point out that errors of measurement are likely to suppress both the consistency and necessity ratios; and that the latter is quite sensitive to the number of failures on the test.

> If an unusually easy behavior is being analyzed, sufficient numbers of failures may not be found and the [necessity] ratio will be low reflecting only chance scores. Thus the necessity of the mastery of the lower set to the acquisition of the terminal behavior cannot be demonstrated. (Capie & Jones, 1971, p. 140)

Consequently, the ratios are also sensitive to low reliability and validity indices. White (1974c) also points out that the ratios lack generalizability; and that high ratios suggest invalid hierarchical relationships. White's first conclusion is that the ratios are more nearly a measure of the degree of difficulty of the content for the examinees. "Their uselessness of hierarchy validation is illustrated by the fact that only five of the 31 connections in three of Gagné's studies [1961, 1962, 1965] meet the condition that all three indexes should be 0.85 or greater, and four of these five involve lower skills which were achieved by all Ss" (p. 63).

2.2.6 Bart, Airasion and Krus

There have been several contributors to the ordering-theoretic approach to hierarchy validation, including Airasion and Bart (1973, 1975), Bart and Airasion (1974),

Bart and Krus (1973), and more recently Baker and Hubert (1977).

Bart and Krus (1973) present a variation on a theme detailed by Gagné (1970a). First, let us assume that all items are dichotomously scored: "0" for the incorrect response, "1" for the correct response. We can then define a "prerequisite" relationship as follows: ". . . success on item *i* is a prerequisite to success on item *j* if and only if the response pattern (0) for items *i* and *j* respectively does not occur" (Bart and Krus, 1973, p. 292). Using the 2x2 response matrix of Figure 1, cells (00), (10), and (11) are defined as *confirmatory* response patterns; while the (01) cell is defined as *disconfirmatory*. For any given set of *n* items it is possible to construct an *nxn* matrix where entries represent the percentage of *disconfirmatory* response patterns for each of the two-item sets. It is then necessary to set a tolerance level for the percentage of disconfirmatory responses above which a relationship between the item-pairs will be rejected. So, for example, if the researcher has a pre-established tolerance level of 5%, and items *i* and *j* have a 4% disconfirmatory response pattern, then the conclusion is simply that item *i* is a prerequisite to item *j*. Naturally, low tolerance levels are recommended so that spurious relationships will not be validated.

More recently Baker and Hubert (1977) have elaborated on the ordering-theoretic model introducing a *directed-graph* representation of the data, and suggesting a statistical procedure for testing the goodness-of-fit of the data to an hypothesized hierarchy. The Baker modification is based on four assumptions:

1. Transitivity applies to *equivalent* item-pairs. That is, two items, i and j, which are both *equivalent* to a third item k, are also *equivalent* to each other. Stated symbolically, if $i \leftrightarrow k$ and $j \leftrightarrow k$, then $i \leftrightarrow j$;

2. Transitivity applies to *prerequisite* item-pairs. That is, two items, i and j, that are both prerequisite to a third item k, are also prerequisite to each other. Stated symbolically, if $i \rightarrow k$ and $j \rightarrow k$ then $i \rightarrow j$;

3. No cycle exists in the directed-graph except among those nodes that are all mutually equivalent;

4. Two equivalent items are prerequisite to the same set of other items.

Not all data-sets will satisfy each of the above assumptions as Baker demonstrates with some simple examples. However, in general, they do obtain under most circumstances.

The goodness-of-fit test is one of comparing all n! possible permutations of the empirically-generated prerequisite matrix. Of course, this can be an enormous task. There are, however, approximations which yield

fairly accurate estimates. Baker suggests using the Cantelli value of $1/(Z^2 + 1)$ for (see Rao, 1965, p. 117) a conservative significance level of the probability of a non-random correspondence between data and theory. The Cantelli inequality simply states that the probability of observing a r value larger than $Z = (r-\Sigma(r)/\sqrt{V(r)}$ is $1/(Z^2 + 1)$, where r is defined as the number of correspondences occurring between the empirical and hypothesized matrices.

There are at least two major limitations to the ordering-theoretic model. First, it is quite obviously limited to pairwise analysis. Consequently the model more easily accommodates itself to linear rather than branching hierarchies. Secondly, the relationships established are function of Σ, a pre-established tolerance limit. This is the way in which the model accounts for errors of measurement.

2.2.7 White

White has been one of the sharpest critics of the various models for validating learning hierarchies (cf. White, 1973, 1974a, 1974b, 1974c; White & Clark, 1973). White (1974a) lists five weaknesses inherent in the previously discussed models which limit their usefulness:

> 1. The elements that comprised the hierarchy were often loosly defined, so that it was possible for someone to

possess one attribute of the element but not another. This led to uncertainty about whether the person could be said to possess the element or not.

2. Often only one question was used for each element to test whether Ss had learned it or not. This meant that no estimate could be made of the prevalence of chance errors or successes, and hence it was not possible to tell whether Ss who answered correctly the question for the higher element but incorrectly the question for the lower element only appeared to have behaved in a way contrary to that required by the hierarchy when they did not really do so.

3. The studies lacked a proper index that could be used to decide whether connections between pairs of elements could be accepted as hierarchical or not. White (1974a) has discussed the shortcomings of the various indexes that were tried.

4. In some studies the elements of the hierarchy were taught to a group of Ss, who were tested on all the elements together after the teaching was completed. In other studies there was no teaching and the Ss were only tested on their possession of the elements. From results obtained by Gagné and Bassler (1963), which showed that although a set of elements may have to be learned hierarchically they can be forgotten in any order, it can be deduced that under either of these procedures it is probable that postulated connections between elements may be wrongly accepted or rejected.

5. In a few studies a small number of Ss was used, which meant that quite a substantial proportion of people in the population from which the Ss were drawn could behave in ways contrary to that required by a valid hierarchy and yet remain undiscovered through not being drawn in the sample. (p. 121ff.)

In order to correct for these five weaknesses, White suggests nine steps that should be implemented in validating a hierarchy.

1. Define the hypothesized highest-order skill in the hierarchy in behavioral terms.

2. Define the remaining lower-order skills by asking Gagné's question vis-a-vis each lower-order skill in turn.

3. Content-validate the hypothesized hierarchy by soliciting the opinion of "experts."

4. Postulate sub-divisions of each skill in the hierarchy so that skill definitions obtain a high degree of precision.

5. Test empirically whether or not these new subdivisions really do represent unique skills.

6. Develop an instructional unit for teaching skill acquisition, embedding mastery tests in it for each skill.

7. Use a minimum of 150 examinees allowing them to work through the instructional unit and to be tested as appropriate.

8. Using frequency counts in a 2x2 matrix, examine the results for large numbers in the (01) cell which would indicate rejection of the postulated connection.

9. Modify the hierarchy to include only those connections that survived the validation process.

White's model is a carefully thought-out process that addresses the major weaknesses found in other models, viz., ill-defined elements, errors of measurement, statistical tests of goodness-of-fit, forgetting and

guessing parameters, and small numbers of examinees. On the other hand, an experimental model such as this is quite difficult to implement within the constraints the researcher usually finds in conducting field research.

One aspect of the model requiring further elaboration is that of the test of inclusion (White & Clark, 1973). This is the only model that can accommodate multi-item data-sets for each objective in the hierarchy and, as a result, can estimate the probability of a randomly selected examinee having answered zero, or one or more times correctly for any objective-pair. White and Clark develop an example for the two-item and three-item case.

Assume a matrix of response frequencies as indicated in Figure 3.

Skill II

Skill I: Questions Correct	0	1	2
2	P_{20}	P_{21}	P_{22}
1	P_{10}	P_{11}	P_{12}
0	P_{00}	P_{01}	P_{02}

Figure 3. A 3x3 response matrix for the two-item case.

White and Clark define the population sub-groups as follows:

P_0 = proportion of the population having neither skill

P_B = proportion of the population having both skills

P_I = proportion of population having only Skill I

P_{II} = proportion of population having only Skill II

Likewise, the conditional probabilities of members of the group answering the appropriate number of items correctly are:

θ_a = probability of examinee with Skill I answering correctly any item for Skill I

θ_b = probability of examinee without Skill I answering correctly any item for Skill I

θ_c = probability of examinee with Skill II answering correctly any item for Skill II

θ_d = probability of examinee without Skill II answering correctly any item for Skill II

For each cell in the matrix the total probability is defined as the product of the probability of the examinee's being in each group and the conditional probability of members of the group answering the appropriate number of items correctly. A probability estimate for each cell in Figure 3 reflecting the nine possible outcomes could be derived if the values of the various Ps and θs were known. For each cell in the matrix estimates of P, the

proportion of the population who have neither skill, both skills, or one or the other skill, as well as θ, the conditional probabilities, can be derived either from the cell frequency using a maximum likelihood procedure or from the marginal totals. In the two-item case P_{02} can be found by substituting the estimates of P and θ in

$$P_{02} = P_0(1-\theta_b)^2\theta_d^2 + P_I(1-\theta_a)^2\theta_d^2 + P_{II}(1-\theta_b)^2\theta_c^2 + P_B(1-\theta_a)^2\theta_c^2$$

Estimates could be calculated for the remaining eight cells, allowing for the generation of a multinomial distribution, and the test of significance could look for deviations outside the observed distribution. However, all this is not necessary according to White and Clark. They suggest that the P_{02} cell would sustain the greatest proportional change if shifts in frequency obtained in the other eight cells. Therefore, it is merely necessary to set a critical value, C, beyond which H_o will be rejected if the observed frequency, f_{02}, in the (02) cell is in excess of C. While White and Clark do not suggest real values for C they do point out that the magnitude of C is directly proportional to a large N and a correspondingly high value of P_{II}.

2.2.8 Dayton and Macready

Dayton and Macready have developed three models for hierarchy specification and validation (Dayton & Macready,

1976a, 1976b; Macready, 1975; Macready & Dayton, 1977; Macready & Mervin, 1973). Model I is a classification (mastery/non-mastery) model with error probabilities associated with each individual item. For a task with n items, the true score patterns are assumed to be 00 . . . 0 and 11 . . . 1, where α_i is the guessing parameter for item i and β_i is the forgetting parameter for item i. Consequently the probability of the j^{th} observed response pattern on an n-item test is:

$$P(j) = P(j|\bar{M})\ P(\bar{M}) + P(j|M)\ P(M)$$

where M = masters

and $\bar{M}$ = non-masters

$$= \left[\prod_{i=1}^{n} \alpha_i^{a_{ij}} (1-\alpha_i)^{1-a_{ij}}\right] \bar{\theta} + \left[\prod_{i=1}^{n} \beta_i^{1-a_{ij}} (1-\beta_i)^{a_{ij}}\right] \theta$$

where $a_{ij} = \{0,1\}$ is the score on the i^{th} item for the j^{th} response.

The computer program, MODEL 3G, developed by the authors (1976b), requires the frequencies for all 2^n possible configurations. Maximum likelihood estimates of the 2n+1 independent parameters are obtained using the Newton-Raphson iteration procedures; chi-square goodness-of-fit tests are also performed. One of the constraints of the program is that the number of items must be greater than four, but less than ten. The lower bound is necessary

since perfect fit of the model to any set of data is always possible due to the number of parameters in the model if the number is less than four.

Model II is also a classification model wherein error probabilities are constant across items. That is, $\alpha_i = \alpha$ and $\beta_i = \beta$. Model II requires the score frequencies for 0, 1 . . . n for an n-item task. The probability of a score of j occurring is given by,

$$P(j) = P(j|\bar{M})\, P(\bar{M}) + P(j|M)\, P(M)$$

$$= \left[\alpha^{s_j} (1-\alpha)^{n-s_j}\right] \bar{\theta} + \left[\beta^{n-s_j} (1-\beta)^{s_j}\right] \theta$$

where s_j = number of correct responses.

Under the assumptions of Model II, the computer program, MODEL 3, will generate the maximum likelihood estimates of the model parameters which number three because of the equal-error assumption. This model is capable of handling up to 100 items, with the minimum being three.

Model III is a generalized version with constant guessing and forgetting parameters which may be written as,

$$P(u) = \sum_{j=1}^{q} P(u|v_j) \cdot \theta_j$$

and $$P(u|v_j) = \prod_{i=1}^{n} \alpha_i^{a_{ij}} (1-\alpha_i)^{b_{ij}} \beta_i^{c_{ij}} (1-\beta_i)^{d_{ij}}$$

The following definitions obtain:

$\mathbf{u}$ = column vector representing the examinee's response pattern, comprised of 0's and 1's.

v = pattern vector which, as a set, represents the response pattern for the a priori hierarchy, comprised of 0's and 2's.

g_{ij} = vector difference, $v_j - u$

a_{ij} = 1 if $g_{ij} = -1$; 0 otherwise. (Examinee guessed)

b_{ij} = 1 if $g_{ij} = 0$; 0 otherwise. (Examinee didn't know item and responded incorrectly)

c_{ij} = 1 if $g_{ij} = 2$; 0 otherwise. (Examinee forgot)

d_{ij} = 1 if $g_{ij} = 1$; 0 otherwise. (Examinee knew item and responded correctly)

For any given element, three of the above will equal zero, the fourth will assume the appropriate value: 1, -1, 0 or 2. The model allows for the estimation of 2^n-1 independent parameters. However, only 2^n-2 may be fitted to any model in order to provide for the goodness-of-fit test. Once the parameters have been estimated, the expected proportions for each 2^n data vectors can be computed. At this point either the chi-square test or likelihood ratios is appropriate for estimating goodness-of-fit.

A computer program, MODEL 5, has been developed by the authors for computing all parameter estimates and testing for significance. The number of items is limited to less than ten, but more than three with no more than

one item per task. In general, both Dayton and Macready Models as well as the White and Clark Model represent a significant advance in the technology of hierarchy specification and validation. Both of the models, because they are probabilistic rather than deterministic, allow for the estimation of errors of measurement, and therefore, for statistical tests of fit. The Dayton and Macready Model also allows for specific errors due to guessing and forgetting. Both methods have advantages and disadvantages that are worth noting. Table 2 summarizes the results of such a comparison. Hambleton and Eignor (1978) have discussed some of these in a recent manuscript.

2.2.9 Summary

This review has presented the critical aspects of eight methodologies employed in the specification and validation of learning hierarchies. Each model varies considerably in its technological sophistication for dealing with the inherent problems of hierarchy validation research. Some of the earlier models for example have considerable intuitive appeal because of their simplicity, while the later models capitalize on recent advances in computer technology. The "state of the art" is far from having reached closure on the issue. Only continued research in the area will produce the desired results,

Table 2

Comparison of Two Methodologies for Validating Hierarchies:

White & Clark vs. Dayton & Macready

Model	Advantages	Disadvantages
White & Clark	Uses multi-item measures for each objective	Valid connections may be rejected if sample size too large
		Tests pairwise connections only
		Probability estimates not maximum likli-hood
Dayton & Macready:	Tests whole hierarchy	Limited to one item per objective
Models I and II	Maximum liklihood estimates	
	χ^2 goodness-of-fit	
	Allows for arbitrary hierarchies, both linear and branching	
	Allows for mastery/non-mastery decisions	
	Suggests hierarchy revisions as well as rejections	
Model III	Tests whole hierarchy	Limited to one item per objective
	Maximum liklihood estimates	
	χ^2 goodness-of-fit	
	Allows for arbitrary hierarchies, both linear and branching	
	Suggests hierarchy revisions as well as rejections	

viz., a straightforward, perhaps eclectic, model which minimizes the weaknesses and maximizes the strengths of each of the preceding methodologies.

CHAPTER III

METHODOLOGY

3.1 Statement of the Problem

The purposes of this study were threefold:

1. To establish using an empirical approach, whether or not a hierarchy exists among selected reading skills;
2. To establish the direction of the hierarchy and assess the strength of the relationship among the component reading skills;
3. To compare several empirical methodologies for establishing hierarchical relationships.

Eight phonics skills and eight structural analysis skills were selected from the test battery, *The Reading Skills Inventory: A Criterion-Referenced Assessment* (Hambleton, 1975). Although each skill had already been assigned by a group of teachers to one of the test levels in the *Inventory*, that assignment was not made exclusively on the basis of perceived hierarchical relationships among the skills. Therefore, the opinions of several content specialists in the field of reading were solicited in order to establish the hypothesized hierarchical relationships. Given an a priori hierarchy, therefore, it then

became possible to apply one or more methodologies for validating the connections.

The selection of these particular sixteen skills was in part based on the author's perusal of several of the major reading programs currently in use. Some of those reviewed include Croft (1976), Wisconsin Design Reading Program, Fountain Valley Program, the SRA Program, and those of major textbook publishers including Ginn, Holt, Scott-Foresman, and Houghton-Mifflin. The basal programs and reading management programs vary considerably in their approach to the teaching of reading. For example, the Croft program uses primarily a skills approach for both word attack and comprehension; whereas, the Harper Row or Lippincott series reflect a more psycholinguistic approach. However, in almost all cases, while the approach may vary, there are fundamental instructional objectives that are common to all methodologies. It is thought that the sixteen skills selected as part of this study are fundamental and common to each series surveyed. Support for this assumption also comes from the Reading Skills Checklist (Larrivee, 1977). The Checklist was developed based on Larrivee's review of the ten most frequently used reading programs in American elementary schools. That checklist includes each of the selected skills of this study.

3.1.1 Theoretical Base

There is a growing body of literature in the area of learning hierarchy research. Gagné, almost two decades ago, stimulated the first investigations with his work in mathematics hierarchies. Since then, there have been numerous contributions to the field, each of which has specified and/or validated hierarchies in specific content areas such as mathematics or science. The basic assumption of all this research is that, notwithstanding individual differences among learners, learning occurs in a sequentially ordered fashion such that prior acquisition of certain skills has a positive impact on the posterior acquisition of other skills. If the results of hierarchy research are generalizable to the learning process as a whole, irrespective of content, it should be possible to specify and validate a hierarchy in the area of reading. Specifically, the theory of hierarchical learning together with the assumption derived from validation methodologies should predict:

1. The ordering of the selected reading skills;
2. The relative degree of dependence among those skills appearing in the hierarchy;
3. The identification of skills which are independent of the hierarchy;
4. The invalid connections in the a priori ordering;

5. The goodness-of-fit estimates;
6. The revisions necessary in the postulated hierarchy.

3.1.2 Hypotheses Tested

The hypotheses tested in this study were as follows:

1. Given that there are 28 possible connections for each of the eight-component sub-hierarchies, and 120 possible connections for the 16-component hierarchy as a whole, then, utilizing the White and Clark (1973) procedure for each component pair,

$$H_0: \quad f_{on} \leq C$$

$$H_1: \quad f_{on} > C$$

where C = pre-established threshold, and f_{on} = the observed frequency in the O2 or O3 cell.

2. Given that there are nine true score response patterns for each of the eight-component sub-hierarchies, and 17 true score patterns for the hierarchy as a whole, then, utilizing the Dayton and Macready Model III (1976a, 1976b), for each true score pattern,

$$H_0: \quad P_{j_{obs}} = P_{j_{pred}}$$

$$H_1: \quad P_{j_{obs}} \neq P_{j_{pred}}$$

Furthermore, the following informal hypotheses were tested as part of this study:

3. That the Dayton and Macready (1976) procedure is preferred to the White and Clark (1973) methodology when the sample size is sufficiently large.

4. That averaging across milti-item sets is superior to a single-item estimation in establishing response patterns.

5. That maximum likelihood estimates of probability utilized in the Dayton and Macready procedure are superior to those employed by White and Clark.

3.2 Methodology

3.2.1 Design

Table 3 specifies the 16 phonics and structural analysis skills selected for this study and measured by the *Inventory*. The phonics and structural analysis sub-hierarchies were considered separately as well as by clusters within each sub-hierarchy. Clustering was necessary because the selected objectives were measured across four different test levels while examinees were only given any two contiguous levels. Specifically, among test levels 2, 4, 5 and 6 an examinee would normally have been administered 2 and 4, 4 and 5, or 5 and 6; but not 4, 5 and 6. Therefore, although a single content-cluster such as structural analysis was measured in levels 4, 5 and 6 it was necessary to data-analyze 4 and 5 separately from 5

Table 3

Sixteen Initially Selected Reading Skills Utilized in the Study

Objective Code	Objective	No. of Items
22	Beginning consonant sounds	6
24	Auditory discrimination: rhyming	5
25	Ending consonant sounds	5
41	Beginning consonant digraphs	9
43	Ending consonant digraphs	8
27	Vowel sounds	6
51	Long/short/r-controlled vowels	15
61	Application vowel principles to nonsense words	8
28	Suffixes denoting syntax	5
45	Inflected/derived from root word	9
55	Prefixes, suffixes	8
56	Root word + affix	5
48	Syllabication by vowel sound	8
63	Syllabication of nonsense words	6
64	Suffixes and syntax (verbs)	5
65	Suffixes and syntax (nouns)	5

and 6 since the logistics of the testing design precluded the tri-level analysis. Because of the clusters six distinct data sets resulted. However, three of these were eliminated from the final comparisons because they contained three or less objective-pairs. The remaining three data sets are displayed in Table 4.

Under the assumptions of the White and Clark model, tests of pairwise connections were conducted. This resulted in the testing of 31 connections: 15, 6 and 10 for Data Sets I, IV and V respectively. The computer program INCLU was written to randomly select the items for each of the pairwise comparisons. Several passes were made through the program for both the two-item as well as the three-item case.

Since the Dayton and Macready Model tests the hierarchy, sub-hierarchy or cluster as a unit, there were 16, 7 and 11 true score response patterns for each of the three clusters. Because the model can only accommodate single-item estimates it was necessary to either eliminate some items from the analysis or to average across items in order to dichotomously score each component. The author felt it desirable to elect the latter procedure since "averaging" would probably yield a more reliable estimate of the examinees' performance on any single objective. Two arbitrary cutting scores were selected to define a "pass":

Table 4

Thirteen Finally Selected Skills Utilized in the Study

Data Set[a]	Objective Code	Objective
I	22	Beginning consonant sounds
(N=2319)	24	Auditory discrimination: rhyming
	25	Ending consonant sounds
	27	Vowel sounds
	41	Beginning consonant digraphs
	43	Ending consonant digraphs
IV	45	Inflected/derived from root word
(N=1118)	48	Syllabication by vowel sound
	55	Prefixes, suffixes
	56	Root word + affix
V	55	Prefixes, suffixes
(N=1688)	56	Root word + affix
	63	Syllabication of nonsense word
	64	Suffixes and syntax (verbs)
	65	Suffixes and syntax (nouns)

[a]Data Sets II, III, and VI were eliminated because of the limited number of objectives in the cluster.

n-1 and n-2, where n equalled the total number of items per objective. Two passes were made through the program, one at each criterion score.

3.2.2 Sample of Examinees

The *Inventory* was administered to the district-wide population of examinees in grades one through six, and a randomly selected sample of 200 examinees in grade seven and eight. The assignment of test level for each examinee was made on the basis of two criteria: (a) the pre-test level which the examinee had taken 12 months earlier, and (b) teacher-judgment. Where possible, the test level assigned to an examinee was that level which the student is known not to have mastered. This judgment was based in large measure on the pre-test data available on each examinee. Where pre-test data were not available, as in the case of an inter-district/intra-district transfer or an absentee, or when, in the best judgment of the pupil's teacher, another level was clearly more appropriate, the test level assignment depended primarily on teacher-judgment. Table 5 displays the test assignments by grade and test level for the grades 1-6 population and the grades 7-8 sample. Examinees in grades seven and eight were selected using a random sampling procedure, stratified by school. Since there were eight middle schools in the district, a sample of twenty-five examinees per school

Table 5

Assignment of Examinees to Test Level by Grade

Grade Level	Test Level					Totals
	PR	I	II	III	IV	
1	950	1536	616	52	11	3165
2	227	1087	1073	292	92	2771
3	50	554	763	633	423	2423
4	37	310	602	744	751	2444
5	5	127	294	638	902	1966
6	3	40	143	573	946	1705
7	--	3	14	28	10	55
8	--	--	1	4	1	6
Totals	1272	3657	3506	2964	3136	14,535

NOTE. The totals of the grades seven and eight rows do not equal 200 (the random sample) because some grade seven and eight examinees were administered a "mature" level of the test specifically designed to test low level skills with older pupils. Response data from the "mature" version was not considered in the study.

was selected. These examinees were assigned to their test level using the same criteria as the lower grade population: (a) pre-test data where available; and (b) teacher-judgment in some cases.

3.2.3 Experimental Controls

Since the data were collected from an actual testing occasion, the controls employed were those normally operant in any well-administered district testing program. Although threats to internal validity were of no concern, threats to external validity could play a significant role in severely limiting the generalizability of the results. A factor of particular interest to the author was the interaction effects of selection biases and test results. Because the entire population in grades one to six was tested the primary source of bias flowed from: (a) elimination of selected examinees because of extenuating circumstances, and (b) the number of absentees on the days of the test administration. In this particular district the following categories of students are systematically eliminated from any general testing procedures:

1. Pupils whose language dominance is other than English. The rule-of-thumb used to measure current language dominance is the length of time the pupil has been a resident in the country. If the pupil has been in the country in excess of one full academic year then he/she

is considered to have sufficient fluency to be tested with the regular school population.

2. Pupils whose current placement is in a special education self-contained unit. Students whose former placement was special education, or whose current status is that of a mainstreamed pupil are not eliminated from general testing occasions. Physically handicapped are tested under whatever supportive conditions necessary to effect valid individual results.

The number of examinees absent on the days of the test administration is somewhat more difficult to estimate. Each examinee was to take two word attack levels beyond the level where mastery was indicated on the pretest. For example, if a pupil on the pre-test showed mastery of Level I, then on this testing occasion the pupil would have been assigned Levels II and III. However, in the case where the pupil had mastered Level III on the pre-test, the only post-test level on which to check mastery was Level IV. Therefore, in some instances pupils were assigned only one level instead of two.

One way of estimating the percent of student participation is to compare the number of students tested with the number enrolled. Table 6 displays the data for the district population in grades one through six. For grades 1, 2, 3 and 4 the assumption of two test levels

per examinee seems warranted. Beyond grade four, however, it is not; and there is no possible way to estimate the percent of participation.

For grades seven and eight, 178 examinees from the random sample of 200, or 89%, participated in the testing. Examination of the participation data indicates that there was possibly a slight selection bias operating but not of sufficient degree to distort the data in any significant way.

Table 6

Comparison of the Number of Examinees with the Number of Enrollees Across Grades

Grade	Number of Examinees[a]	Number of Enrollees	Estimated Percent[b]
1	3165	1758	100%
2	2771	1570	100
3	2423	1437	100
4	2444	1557	100
5	1966	1632	--[c]
6	1705	1531	--

[a]This may be a duplicate count since each examinee could take two test levels.

[b]This percent is estimated based on the assumption of two test levels per enrollee.

[c]At this grade level the assumption of two test levels per enrollee was not warranted.

3.2.4 Instrumentation

The Reading Skills Inventory: A Criterion-Referenced Assessment was designed to assist the implementation of a program of individually-paced reading instruction in grades K through 8. One component of the *Inventory* consists of a set of criterion-referenced word-attack skills measures at each of five instructional levels: pre-reading, Level I, II, III, and IV. A second component consists of nine levels of reading comprehension measures. However, in this study only results from the Word Attack component were analyzed. The original test specifications were generated based on the district's reading curriculum scope and sequence. Since the *Inventory* was intended to be a survey criterion-referenced measure rather than a diagnostic instrument, only selected objectives in the curriculum were included for measurement. The decision regarding the inclusion or exclusion of a particular objective was made primarily by judgment sampling based on the professional opinion of reading specialists and teachers in the district. The primary purpose to which the test results would be directed was to be the tracking of individual student performance. Secondarily the test results would also determine ad hoc classroom groupings, as well as influence decisions relative to student pacing. Table 7 displays the content outline of each Work Attack

Table 7

Content and Item Outline of the *Reading Skills Inventory* by Level

Curriculum Code	Obj. Code	Objective	Item Numbers	Total No. Items
		Pre-Reading		
VPM-01		Seriation by size	01-04	4
VPM-02		Classification by function	05-07	3
VPM-08		Letter & number completion	08-16	9
VPM-09		Multiple letter matching	17-22	6
VPM-10		Single letter matching	23-26	4
VPM-11		Word matching	27-30	4
LOL-04		Oral vocabulary: home content	31-33	3
LOL-07		Oral vocabulary: social content	34-36	3
PRR-01		Letter-sound association	37-44	8
PRR-03		Initial sounds	45-47	3
PRR-05		Visual discrimination: letter patterns	48-53	6
		Word Attack Level I		
PRR-02		Letter production	01-04	4
* PH-01	22	Beginning consonant sounds (BCS)	05-10	6
PH-02		Picture-BCS association	11-13	3
* PH-03	24	Auditory discrimination: rhyming	14-18	5
* PH-05	25	Ending consonant sounds	19-23	5
PH-06		Position of consonant sound	24-32	9
* PH-10	27	Vowel sounds	33-38	6
SA-01		Suffixes denoting syntax	39-43	5
DS-01		Seriation by alphabet	44-48	5
PH-04		Context clues	49-53	5
		Word Attack Level II		
* PH-07	41	Beginning consonant digraphs	01-09	9
PH-08		Rhyming	10-14	5
* PH-09	43	Ending consonant digraphs	15-22	8
PH-11		Long/short vowel sounds	23-31	9
* SA-02	45	Inflected/derived from root word	32-40	9

Table 7 (continued)

Curriculum Code	Obj. Code	Objective	Item Numbers	Total No. Items
SA-03		Compound words	41-48	8
SA-04		Contractions	49-56	8
* SA-07	48	Syllabication by vowel sound	57-64	8
DS-02		Alphabetize by first letter	65-69	5
		Word Attack Level III		
PH-12		Long/short/r-controlled vowels	01-15	15
PH-14		Word-vowel sound association	25-32	8
PH-15		Application of vowel sound principles	33-40	8
* SA-05	55	Prefixes, suffixes	41-48	8
* SA-06	56	Root word + affix	49-53	5
DS-03		Alphabetize by 2nd and 3rd letters	54-58	5
		Word Attack Level IV		
PH-16		Application of vowel principles: nonsense words	01-08	8
SA-08		Recognition of initial vowel sounds	09-16	8
* SA-09	63	Syllabication of nonsense words	17-22	6
* SA-11	64	Suffixes and syntax (verbs)	23-27	5
* SA-12	65	Suffixes and syntax (nouns)	28-32	5
DS-04		Dictionary guide words	33-37	5
DS-06		Dictionary meaning and context	38-42	5
SA-10		Articulation of nonsense words	43-48	6
		Comprehension: All Levels		
RC-01		Main idea	3/level	
RC-02		Stated detail	3/level	
RC-03		Inferred detail	3/level	
RC-04		Inference	3/level	
RC-05		Vocabulary	3/level	

NOTE: The asterisked objectives are those finally selected for inclusion in the study.

component as well as the number of items utilized to measure each objective. For purposes of tabulation an abbreviated form of the objective has been listed. However, the reader will find the complete objective as stated for the original test specifications in Appendix A.

A validation study conducted by the author of the *Inventory* (Hambleton, 1975) established the adequacy of each level in the battery. Validation data included both an analysis of item difficulties and item discrimination indices. Reliability assessment included measures of internal consistency; parallel form reliability; and proportion of agreement of mastery decisions based on two performance standards (80% and 100%). In addition a content validation study was conducted.

3.2.5 Procedures

The *Inventory* was administered over a three week period in the Spring. Most items on the test could be group-administered. However, some objectives required a one-to-one administration. These were given by the regular classroom teacher as time allowed during the testing period. Test results were then sent to a service bureau for hand or machine scoring. All items were scored dichotomously: "0" for a fail, and "1" for a pass. The district had specified arbitrary cutting scores for

each objective in order to determine mastery on any given objective. However, these were not considered in this study. Rather, for each objective, a cutting score of either n-1 or n-2 was used. This was particularly appropriate in implementing the Dayton and Macready model. For the White and Clark model the item scores were selected on a random basis.

In order to establish an a priori hierarchy based on the judgment of experts in the field of reading a sample of 23 content specialists was asked to respond to a pairwise comparison task. The backgrounds of the specialists included both special as well as regular education; eight were university affiliated, nine were elementary teachers of reading, four were administrators of reading programs and two were doctoral candidates in the field of reading. Of the 23 respondents, two were eliminated from the final sample because of their failure to understand the task.

Each respondent was asked to examine the 56 pairs of objectives resulting from the two 8-objective clusters: one cluster of phonics skills and a second of structural analysis skills. The ordering of objectives within each pair was identical for all respondents, but the order of presentation of the pairs was random. The respondent was asked to evaluate for each objective-pair whether mastery of one objective of the pair was necessary prior to the

mastery of the second objective; and if so, which objective should be mastered first. If the respondent felt that the relationship between the objectives was not a necessary one but merely a facilitating one, then the respondent indicated which objective would facilitate the other. Finally, the respondents were allowed the option of indicating no relationship between the objectives; that is, each objective could be mastered independently of the other.

In order to insure that all respondents clearly understood the task, two bogus objectives were included in the sequence, thus increasing the number of pairs evaluated to 72. The two respondents eliminated from the sample failed to correctly classify the bogus objectives in the third option, i.e., these objectives could clearly be mastered independent of all other objectives in the cluster. A copy of the hierarchy specification forms used to complete this task is included in Appendix B.

3.2.6 Computer Programs

The author is grateful to George M. Macready of the University of Maryland for providing a listing of the Dayton and Macready program, MODEL 5, which was used to conduct that segment of the data analysis. No substantive changes were made in the program, only those necessary for compatability with the CDC system.

It was necessary to write a program, INCLU, based on the White and Clark model. The author would like to thank Richard Rovinelli of the American Board of Family Practice, for providing the initial version of that program and Leah Hutten of the University of Massachusetts for her work in revising and modifying the program to meet the data analysis requirements of this study. The program allows for either the two-item or three-item pair-wise comparison, with a sub-routine for randomly selecting items for each comparison, and accepts scored item data from the input data set.

C H A P T E R I V

RESULTS

4.1 Comparison Across Data Sets

In this section the data are presented for proposing the hierarchies resulting from the three distinct methodologies applied in this research study. No attempt will be made in this section to compare and contrast various approaches, nor will any reference to the content of the objectives be made. The content significance of the results as well as a comparative analysis across methodologies are provided in Section 4.2.

4.1.1 The Expert Judgment Hierarchies

One of the primary purposes of this study was to establish an a priori ordering of the objectives based on the collective judgment of professionals in the field of reading. The results of such a hierarchy specification task are displayed in Tables 8, 9 and 10 for Data Sets I, IV and V, respectively.

In Data Set I only three objective-pair are clearly viewed by the majority of respondents as having independent components. However, although all other pairs are viewed as having a relationship between component

Table 8

Distribution of Responses to Hierarchy Specification Task for Data Set I (N=21)

Objective Pair Code	Responses					
	Necessary		Facilitates			Order Irrelevant[d]
A-B	A**B[a]	B**A	A*B[b]	B*A	Each[c]	
24-22	4	3	3	1	3	7
22-25	7	0	10	0	1	3
41-22	0	8	1	8	3	1
22-43	6	0	11	0	1	3
27-22	0	3	2	6	1	9
24-25	5	1	7	1	1	6
24-41	3	1	6	0	3	8
24-43	2	1	8	1	2	7
24-27	3	1	10	1	1	5
25-41	3	1	6	0	3	8
43-25	0	6	1	5	5	3
25-27	2	0	3	0	4	12
43-41	2	3	0	6	7	3
41-27	1	0	1	0	2	17
27-43	2	0	1	2	4	12

[a]The A**B notation should be read, "Prior mastery of A is necessary for the mastery of B," and vice versa for B**A.

[b]The A*B notation should be read, "Prior mastery of A facilitates the mastery of B," and vice versa for B*A.

[c]Each facilitates the mastery of the other.

[d]Mastery of A is entirely independent of mastery of B and vice versa.

Table 9

Distribution of Responses to Hierarchy Specification Task for Data Set IV (N=21)

Objective Pair Code	Responses					
	Necessary		Facilitates			Order Irrelevant[d]
A-B	A**B[a]	B**A	A*B[b]	B*A	Each[c]	
45-48	1	1	0	4	1	14
45-55	3	3	1	5	7	2
45-56	3	2	4	2	8	2
48-55	3	0	0	2	0	15
48-56	2	1	0	1	1	16
55-56	5	3	7	1	5	0

[a]The A**B notation should be read, "Prior mastery of A is necessary for the mastery of B," and vice versa for B**A.

[b]The A*B notation should be read, "Prior mastery of A facilitates the mastery of B," and vice versa for B*A.

[c]Each facilitates the mastery of the other.

[d]Mastery of A is entirely independent of mastery of B and vice versa.

Table 10

Distribution of Responses to Hierarchy Specification Task for Data Set V (N=21)

Objective Pair Code	Responses					
	Necessary		Facilitates			Order Irrelevant[d]
A-B	A**B[a]	B**A	A*B[b]	B*A	Each[c]	
55-56	5	3	7	1	5	0
63-55	2	1	2	1	4	11
55-64	8	1	4	3	5	0
65-55	5	4	0	7	5	0
56-63	2	1	0	4	1	13
64-56	2	4	2	5	5	3
56-65	3	1	5	1	8	3
63-64	2	0	2	3	2	12
63-65	2	0	1	3	0	15
64-65	0	2	1	4	7	7

[a]The A**B notation should be read, "Prior mastery of A is necessary for the mastery of B," and vice versa for B**A.

[b]The A*B notation should be read, "Prior mastery of A facilitates the mastery of B," and vice versa for B*A.

[c]Each facilitates the mastery of the other.

[d]Mastery of A is entirely independent of mastery of B and vice versa.

objectives, the strength and direction of the relationship is somewhat obscured by the wide variations in response patterns. This observation holds true for the remaining two data sets.

In Data Set IV more than 50% of the respondents saw only an independent relationship among three of the six possible objective-pair. In Data Set V four of the ten objective-pair were virtually eliminated from the hierarchy by an "order irrelevant" classification by more than 50% of the respondents.

Given the wide variability for the remaining objective-pair, and given the possibility that the hierarchy specification task could have been perceived differently by each of the 21 respondents, it seemed advantageous to collapse objective-pair data across response patterns. Therefore, for purposes of generating a hypothetical hierarchy, all A**B and A*B patterns were collapsed as one response indicating an "A prior to B" relationship; the B**A and the B*A patterns were collapsed as a second response indicating a "B prior to A" relationship. This procedure, while possibly establishing the direction of the relationship, leaves the strength of the relationship undecided. Table 11 summarizes the results for all objective-pair judged by the experts. Clearly, patterns begin to emerge and it is possible to suggest tentative hierarchies among the objective-pair for each data set.

Table 11

Collapsed Responses to Hierarchy Specification Task For All Objective-Pairs Judged by the Experts (N=21)

Objective A-B	Judgment: A prior to B	B prior to A	Each facilitates other	Order Irrelevant
24-22	7	4	3	7
22-25	17	0	1	3
41-22	1	16	3	1
22-43	17	0	1	3
27-22	2	9	1	9
24-25	12	2	1	6
24-41	9	1	3	8
24-43	10	2	2	7
24-27	13	2	1	5
25-41	9	1	3	8
43-25	1	11	5	3
25-27	5	0	4	12
43-41	2	9	7	3
41-27	2	0	2	17
27-43	3	2	4	12
45-48	1	5	1	14
45-55	4	8	7	2
45-56	7	4	8	2
48-55	3	2	0	15
48-56	2	2	1	16
55-56	12	4	5	0
55-56	12	4	5	0
63-55	4	2	4	11
55-64	12	4	5	0
65-55	5	11	5	0
56-63	2	5	1	13
64-56	4	9	5	3
56-65	8	2	8	3
63-64	4	3	2	12
63-65	3	3	0	15
64-65	1	6	7	7

There are six objectives under consideration in Data Set I. From the response data, all six maintain a position within each of the proposed hierarchies. There appears to be some agreement among the experts on the position of objectives 43, 41, and 25. The largest disagreement concerns objectives 27, 24, and 22. There are three possible arrangements and response data to mildly support each of these possibilities. Thus, we have proposed three hierarchies for Data Set I displayed in Figure 4. A full discussion of the educational and practical significance of these hierarchies will be postponed until a later section.

Data Set IV is a much smaller collection of objective-pair, viz., four objectives and six pairs. The results are fairly evident. Objective 48 is judged almost unanimously as not part of the hierarchy. The remaining three objectives, 45, 55, and 56, are placed by the respondents in the hierarchy displayed in Figure 5.

Data Set V consisted of five objectives, one of which was judged to be outside the hierarchy, viz., 63. The remaining four, objectives 55, 56, 64, and 65, were assigned the hierarchical order displayed in Figure 6. In this data set, as with the previous one, there was relatively high agreement among the respondents.

Using the response patterns from Table 11 it was possible to calculate the percent of agreement between

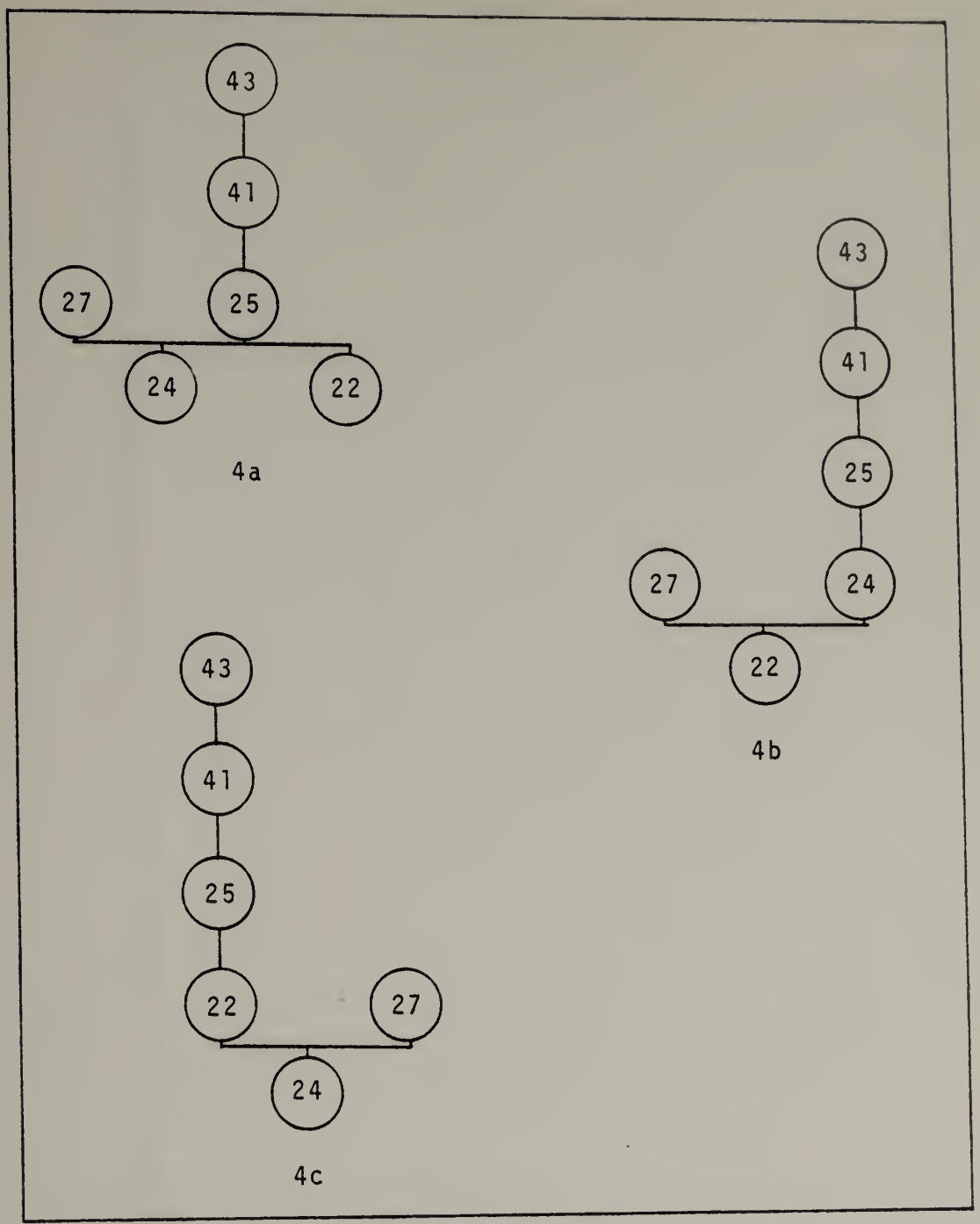

Figure 4. Proposed hierarchy for Data Set I based on response data in Table 11.

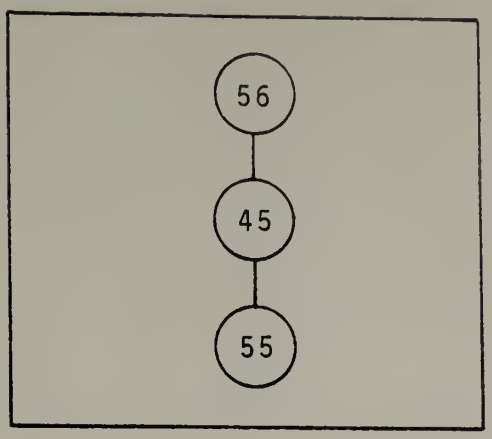

Figure 5. Proposed hierarchy for Data Set IV based on response data in Table 11.

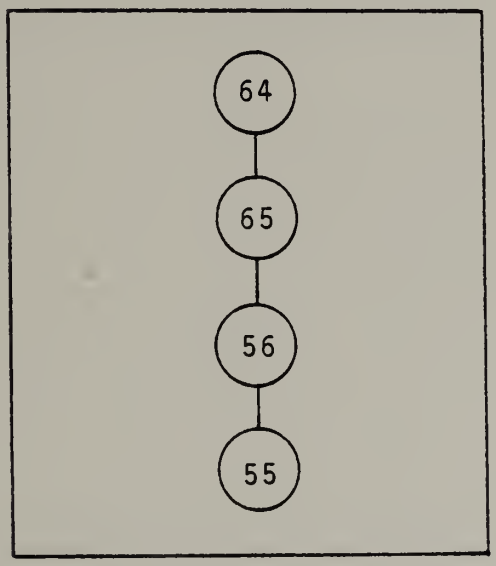

Figure 6. Proposed hierarchy for Data Set V based on response data in Table 11.

the judgment data and the proposed hierarchies of Figures 4, 5, and 6. The index of agreement was derived by estimating the percent of responses favoring the objective's position in the hierarchy based on the total number of judgments for that hierarchy. In Table 12 the reader will note that the indices range from a low of 25% to a high of 64%. However, the average for a given hierarchy is usually greater than 50%. In the case of Data Set I, two of these proposed hierarchies have the same average index of agreement; one is somewhat lower. Data Set V has the lowest average index, 44.4%. The depression of this average index is due to the inclusion of objective 64 in the hierarchy, the position of which enjoys only a 25% agreement with the judgment responses.

4.1.2 The Dayton and Macready Hierarchies

The first of two empirical procedures utilized to validate the hierarchies was that of Dayton and Macready. The criterion score which determined mastery was n-1 on the first pass and n-2 on the second, where n equalled the total number of items measuring a given objective. Table 13 summarizes the maximum likelihood estimates of the parameters and their standard errors at each criterion level for each data set. The frequency data generated from these parameter estimates are presented in Tables 14,

Table 12

Percent of Agreement Between Judgment Data and Proposed Judgment Hierarchies

Data Set	Objective Code	Hierarchy a	Hierarchy b	Hierarchy c
I	43	56.2%	56.2%	56.2%
	41	57.1	57.1	57.1
	25	58.1	58.1	58.1
	24	48.6	38.1	48.6
	27	60.0	52.4	60.0
	22	62.9	60.0	62.9
	Average	57.2	53.7	57.2
IV	56	55.5		
	45	46.0		
	55	55.5		
	Average	52.3		
V	64	25.0		
	65	47.6		
	56	50.0		
	55	54.8		
	Average	44.4		

Table 13

Maximum Likelihood Estimates of Parameters and Their Standard Errors for All Data Sets at Each Criterion Level

	Data Set I				Data Set IV				Data Set V			
	n-1		*n*-2		*n*-1		*n*-2		*n*-1		*n*-2	
Parameter	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE
θ_1[a]	0.016	0.41E-02	0.011	0.303E-02	0.999	0.372E-01	0.999	0.259E-01	0.965	0.358E-01	0.000	0.163E+17
θ_2	0.066	0.776E-02	0.029	0.509E-00	0.000	0.381E-01	0.000	0.268E-01	0.034	0.389E-01	0.000	0.212E+17
θ_3	0.009	0.588E-02	0.000	0.358E-02	0.000	0.263E-01	0.000	0.160E-01	0.000	0.996E-02	0.000	0.212E+17
θ_4	0.043	0.792E-02	0.029	0.582E-02	0.036	0.380E-01	0.000	0.132E-01	0.000	0.490E-02	0.000	0.212E+17
θ_5	0.245	0.148E-01	0.270	0.154E-01	--	--	--	--	0.000	0.252E-02	0.000	0.212E+17
θ_6	0.050	0.120E-01	0.000	0.112E-01	--	--	--	--	--	--	--	--
α[b]	0.257	0.163E-01	0.293	0.189E-01	0.503	0.113E-01	0.654	0.981E-02	0.757	0.586E-02	0.999	0.537E+08
β[c]	0.059	0.313E-02	0.029	0.206E-02	0.999	0.185E+00	0.999	0.186E+04	0.999	0.139E+00	0.999	0.126E+02

[a] θ_n is defined as the proportion at the nth true score pattern.

[b] α is the guessing probability.

[c] β is the forgetting probability.

Table 14

Observed and Predicted Frequencies, Chi-Square and Probability Estimates for all Response Patterns in Data Set I

	n-1 Criterion				n-2 Criterion			
PATTERN	OBS N	PRED N	CHI-SQUARE	PROB	OBS N	PRED N	CHI-SQUARE	PROB
000000	* 10	8.29	.35	.0036	* 1	3.60	1.88	.0016
100000	* 28	35.45	1.56	.0153	* 10	12.75	.59	.0055
010000	2	3.39	.57	.0015	2	1.51	.15	.0007
110000	* 28	20.75	2.53	.0089	* 5	6.12	.20	.0026
001000	3	3.05	.00	.0013	2	1.51	.15	.0007
101000	35	15.29	25.39	.0066	9	6.12	1.35	.0026
011000	3	4.21	.34	.0018	1	1.46	.14	.0006
111000	* 52	55.94	.27	.0241	* 33	30.76	.16	.0133
000100	3	2.92	.00	.0013	3	1.50	1.49	.0006
100100	11	13.21	.37	.0057	7	5.52	.39	.0024
010100	3	2.13	.34	.0009	1	.87	.01	.0004
110100	17	22.65	1.41	.0098	10	10.71	.04	.0046
001100	5	2.02	4.39	.0009	2	.87	1.46	.0004
101100	61	20.76	77.95	.0090	25	10.71	19.02	.0046
011100	8	16.93	4.71	.0073	5	8.79	1.63	.0038
111100	* 269	267.72	.00	.1154	* 298	288.80	.29	.1245
000010	2	2.86	.26	.0012	1	1.49	.16	.0006
100010	5	12.28	4.31	.0053	3	5.28	.98	.0023
010010	0	1.20	1.20	.0005	1	.62	.21	.0003
110010	9	7.63	.24	.0033	5	2.56	2.29	.0011
001010	1	1.08	.00	.0005	0	.62	.62	.0003
101010	9	5.74	1.84	.0025	1	2.56	.95	.0011
011010	0	1.91	1.91	.0008	0	.64	.64	.0003
111010	35	26.72	2.56	.0115	11	13.86	.59	.0060
000110	1	1.03	.00	.0004	0	.62	.62	.0003
100110	2	5.02	1.82	.0022	3	2.32	.19	.0010
010110	6	1.19	19.21	.0005	4	.39	32.98	.0002
110110	12	15.21	.68	.0066	5	5.56	.05	.0024
001110	3	1.15	2.92	.0005	0	.39	.39	.0002
101110	49	14.56	81.38	.0063	20	5.56	37.49	.0024
011110	3	13.24	7.92	.0057	4	4.76	.12	.0021
111110	* 184	211.11	3.48	.0910	* 109	157.43	14.90	.0679
000001	0	2.86	2.86	.0012	0	1.49	1.49	.0006
100001	2	12.26	8.59	.0053	0	5.28	5.28	.0023
010001	1	1.18	.02	.0005	0	.62	.62	.0003
110001	4	7.38	1.55	.0032	0	2.56	2.56	.0011
001001	0	1.06	1.06	.0005	0	.62	.62	.0003
101001	5	5.50	.04	.0024	2	2.56	.12	.0011
011001	1	1.67	.27	.0007	1	.64	.20	.0003
111001	18	22.83	1.02	.0098	14	13.86	.00	.0060
000101	1	1.02	.00	.0004	1	.62	.22	.0003
100101	2	4.78	1.62	.0021	3	2.32	.19	.0010
010101	2	.95	1.13	.0004	0	.39	.39	.0002
110101	6	11.32	2.50	.0049	6	5.56	.03	.0024
001101	1	.91	.00	.0004	0	.39	.39	.0002
101101	26	10.67	21.98	.0046	15	5.56	16.02	.0024
011101	4	9.35	3.06	.0040	7	4.76	1.05	.0021
111101	111	148.71	9.56	.0641	123	157.43	7.53	.0679
000011	1	1.00	.00	.0004	0	.61	.61	.0003
100011	3	4.46	.47	.0019	2	2.22	.02	.0010
010011	0	.63	.63	.0003	0	.29	.29	.0001
110011	7	6.13	.12	.0026	1	2.18	.64	.0009
001011	2	.59	3.33	.0003	1	.29	1.69	.0001
101011	11	5.48	5.53	.0024	5	2.18	3.62	.0009
011011	2	4.16	1.12	.0018	1	1.38	.10	.0006
111011	55	65.42	1.66	.0282	35	43.53	1.67	.0188
000111	4	.57	20.27	.0002	2	.29	10.03	.0001
100111	4	5.23	.29	.0023	0	2.08	2.08	.0009
010111	4	3.91	.00	.0017	4	1.28	5.74	.0006
110111	15	61.45	35.11	.0265	12	40.09	19.68	.0173
001111	3	3.90	.20	.0017	4	1.28	5.74	.0006
101111	97	61.22	20.89	.0264	47	40.09	1.18	.0173
011111	32	60.77	13.62	.0262	35	39.76	.57	.0171
111111	* 1036	974.61	3.86	.4203	* 1417	1340.22	4.39	.5779

15, and 16 for Data Sets I, IV, and V respectively. In all cases the values for the guessing and forgetting parameter were unrestricted. Using the MODEL 5 program, a solution was obtained (convergence to a criterion of .1E-04) for Data Sets I and IV on both passes. For Data Set V the maximum number of iterations (100) was reached without attaining convergence. This solution, therefore, is questionable. The user-specified true score patterns reflected a linear hierarchy in all cases.

Table 17 summarizes the chi-square goodness of fit statistics for each data set at both criterion levels. The significance of the Pearson Chi-Square is applicable, of course, only to the complete data set. And as the reader will note, all are significant. However, inspection of the discrete chi-square variable in Tables 14 through 16 shows several true score patterns having non-significant outcomes. In Data Set I, for example, at the n-1 criterion score all but the last true score pattern (111111) have non-significant chi-square values; at the n-2 criterion score only the last two (111110 and 111111) are significant while the remaining five are not. In Data Set IV only the first pattern (0000) at the n-2 criterion score is non-significant; all others are. In Data Set V three patterns (10000 , 11100 and 11111) at the n-1 criterion score are non-significant; all remaining patterns are. What this suggests is that for the entire data set the

Table 15

Observed and Predicted Frequencies, Chi-Square and Probability Estimates for All Response Patterns in Data Set II

	n-1 Criterion				*n*-2 Criterion			
PATTERN	OBS N	PRED N	CHI-SQUARE	PROB	OBS N	PRED N	CHI-SQUARE	PROB
0000	* 34	48.06	4.11	.0430	* 11	15.92	1.52	.0142
1000	* 131	69.08	55.48	.0618	* 58	30.17	25.65	.0270
0100	21	69.08	33.46	.0618	19	30.17	4.13	.0270
1100	* 200	69.87	242.35	.0625	* 151	57.16	154.01	.0511
0010	0	69.08	69.08	.0618	1	30.17	28.20	.0270
1010	23	69.87	31.44	.0625	19	57.16	25.48	.0511
0110	0	69.87	69.87	.0625	2	57.16	53.23	.0511
1110	* 26	70.66	28.23	.0632	* 43	108.29	39.37	.0969
0001	23	89.32	49.25	.0799	6	30.17	19.37	.0270
1001	128	69.87	48.36	.0625	76	57.16	6.20	.0511
0101	22	69.87	32.79	.0625	13	57.16	34.12	.0511
1101	341	70.66	1034.19	.0632	365	108.29	608.45	.0969
0011	1	69.87	67.88	.0625	1	57.16	55.18	.0511
1011	34	70.66	19.02	.0632	42	108.29	40.58	.0969
0111	3	70.66	64.79	.0632	5	108.29	98.53	.0969
1111	* 131	71.46	49.58	.0639	* 306	205.16	49.55	.1835

* = True Score Pattern

Table 16

Observed and Predicted Frequencies Chi-Square and Probability Estimates for all Response Patterns in Data Set V

	n-1 Criterion				*n*-2 Criterion			
PATTERN	OBS N	PRED N	CHI-SQUARE	PROB	OBS N	PRED N	CHI-SQUARE	PROB
00000	* 9	3.99	6.25	.0024	* 1	1687.98	1685.98	1.0000
10000	* 4	4.29	.01	.0025	* 2	.00	2365.69	.0000
01000	20	4.91	46.32	.0029	3	.00	5325.80	.0000
11000	* 5	13.37	5.24	.0079	* 3	.00	*******	.0000
00100	58	4.91	573.60	.0029	9	.00	*******	.0000
10100	16	13.37	.51	.0079	3	.00	*******	.0000
01100	139	15.31	998.89	.0091	21	.00	*******	.0000
11100	* 53	41.68	3.07	.0247	* 16	.00	*******	.0000
00010	4	4.91	.16	.0029	0	.00	.00	.0000
10010	1	13.37	11.44	.0079	0	.00	.00	.0000
01010	13	15.31	.34	.0091	0	.00	.00	.0000
11010	12	41.68	21.13	.0247	0	.00	.00	.0000
00110	61	15.31	136.27	.0091	23	.00	*******	.0000
10110	29	41.68	3.86	.0247	7	.00	*******	.0000
01110	243	47.73	798.65	.0283	97	.00	*******	.0000
11110	* 162	129.93	7.91	.0770	* 110	.00	*******	.0000
00001	3	4.91	.74	.0029	0	.00	.00	.0000
10001	2	13.37	9.67	.0079	0	.00	.00	.0000
01001	7	15.31	4.51	.0091	1	.00	*******	.0000
11001	2	41.68	37.78	.0247	0	.00	.00	.0000
00101	14	15.31	.11	.0091	8	.00	*******	.0000
10101	8	41.68	27.22	.0247	7	.00	*******	.0000
01101	58	47.73	2.20	.0283	36	.00	*******	.0000
11101	54	129.93	44.38	.0770	36	.00	*******	.0000
00011	3	15.31	9.90	.0091	1	.00	590.41	.0000
10011	4	41.68	34.06	.0247	1	.00	*******	.0000
01011	13	47.73	25.27	.0283	3	.00	*******	.0000
11011	16	129.93	99.98	.0770	8	.00	*******	.0000
00111	43	47.73	.47	.0283	29	.00	*******	.0000
10111	47	129.93	52.93	.0770	69	.00	*******	.0000
01111	201	148.80	18.30	.0882	267	.00	*******	.0000
11111	* 384	405.04	1.09	.2400	* 927	.00	*******	.0000

* = True Score Pattern

Table 17

Chi-Square Estimates for All Data Sets at Each Criterion Level

Statistic	Data Set I		Data Set IV		Data Set V	
	n-1 Criterion	n-2 Criterion	n-1 Criterion	n-2 Criterion	n-1 Criterion	n-2 Criterion
Likelihood ratio	0.352 E+03	0.185 E+03	0.161 E+04	0.120 E+04	0.167 E+04	0.539 E+05
Pearson χ^2	412.56	217.22	1899.9	1243.6	2982.3	----[a]
df	55	55	9	9	24	24

[a]Value of statistic exceeded the limits of the program.

sample size is so large that it would be impossible not to achieve significance. In other words, a large N has led to the rejection of the null hypothesis. Notwithstanding the fact that the compatability between the observed and the theoretical frequency distribution is low, the author would like to consider the value of the discrete chi-square variable for each true score pattern. It then becomes possible to specify tentative hierarchies for some of the data sets. These are displayed in Figure 7.

4.1.3 The White-Clark Hierarchies

The second empirical procedure utilized to validate the hierarchies was that of White and Clark. This procedure is based on the pair-wise comparisons of all objectives in any given sequence utilizing either two or three items per objective to assess mastery. The program INCLU was used to generate the frequencies based on a random selection of two items per objective on the first pass and three items per objective on the second. In all other respects the data sets remained the same except the way in which the response patterns were analyzed. For the two-item case, three replications were generated, while for the three-item case, two replications were generated. Tables 18 through 23 summarize the appropriate statistics for all data sets: the estimated probability

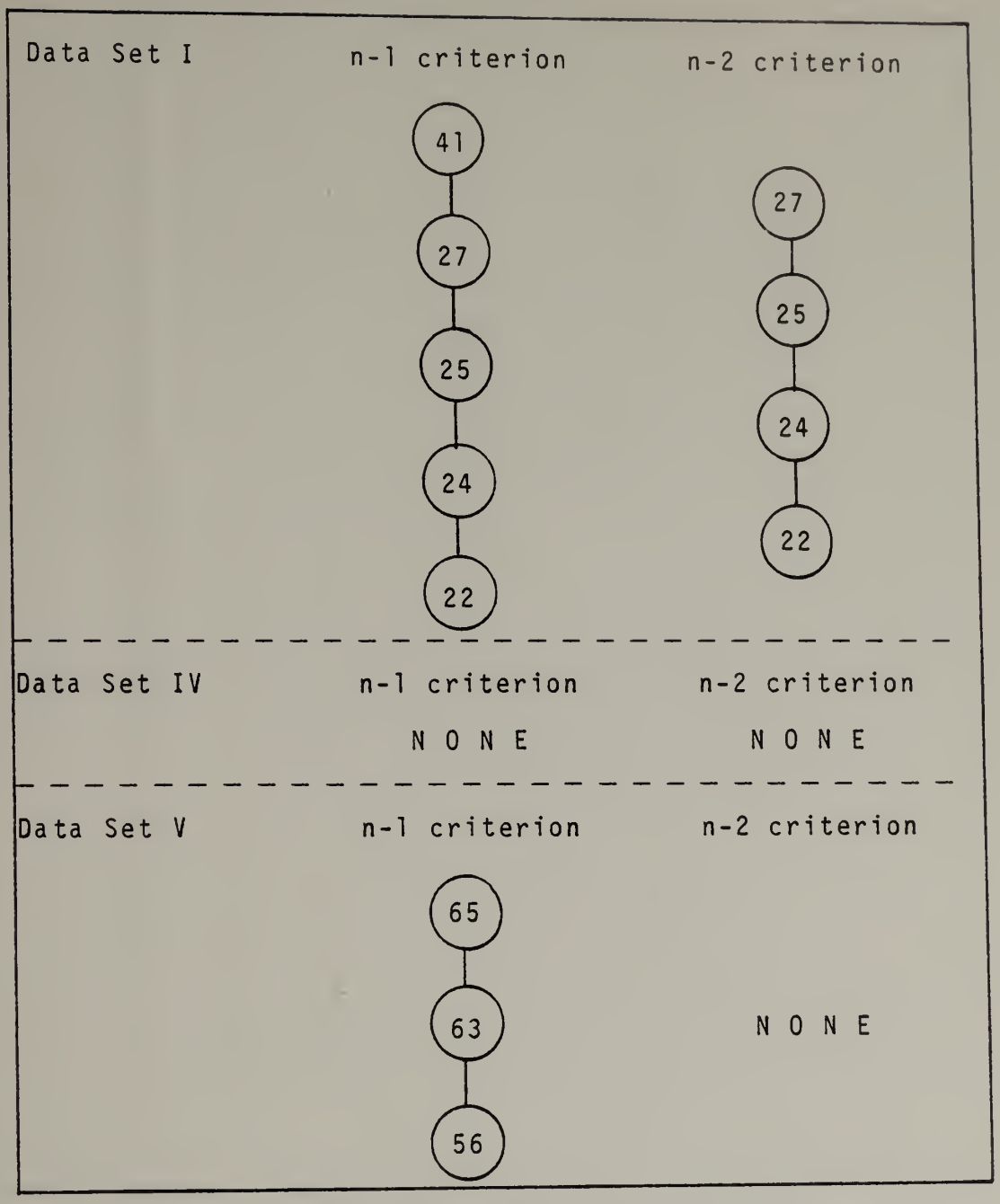

Figure 7. Proposed hierarchies for Data Sets I, IV, and V based on discrete chi-square values of true score patterns found in Tables 13, 14, and 15.

Table 18

Estimated Probability of the 02 Event, Mean, Standard Deviation, Critical Value, and Observed Frequencies in the 02 Cell for Data Set I, Two-Item Case, Over Three Replications (N=2319)

Objective Code	Replication 1					Replication 2					Replication 3				
	P_{02}[a]	$\bar{X}$	SD	C[b]	f_{02}	P_{02}	$\bar{X}$	SD	C	f_{02}	P_{02}	$\bar{X}$	SD	C	f_{02}
25-41	*.01*	0	4.79	14	33	*.01*	0	4.79	14	30	*.01*	0	4.79	14	37
25-43	*.01*	0	4.79	14	34	.01	23	4.79	37	28	*.01*	0	4.79	14	27
27-41	.02	46	6.74	66	0	*.009*	0	4.54	14	21	.01	23	4.79	37	0
27-43	.02	46	6.74	66	66	*.008*	0	4.28	13	19	*.008*	0	4.28	13	20
41-43	.54	1252	24.00	1324	165	.58	1345	23.76	1416	143	.01	23	4.79	37	106
22-24	.01	23	4.79	37	49	.95	2203	10.49	2234	51	.01	23	4.79	37	40
22-25	.01	23	4.79	37	43	.01	23	4.79	37	46	.01	23	4.79	37	41
22-27	.01	23	4.79	37	53	.89	2064	15.07	2109	160	.89	2064	15.07	2109	173
22-41	.75	1739	20.85	1802	65	.71	1646	21.85	1712	139	.77	1786	20.27	1847	51
22-43	.80	1855	19.26	1913	40	.01	23	4.79	37	30	*.01*	0	4.79	14	45
24-25	.02	46	6.74	66	82	.02	46	6.74	66	49	.03	70	8.21	95	72
24-27	.03	70	8.21	95	76	.02	46	6.74	66	63	.03	70	8.21	95	52
24-41	.01	23	4.79	37	63	.01	23	4.79	37	49	.01	23	4.79	37	70
24-43	.01	23	4.79	37	62	.01	23	4.79	37	35	.01	23	4.79	37	42
25-27	.02	46	6.74	66	42	.01	23	4.79	37	42	.01	23	4.79	37	26

[a] P_{02} is estimated from f_{02}/N when INCLU-value equalled 0. Estimated values are in italics.

[b] Critical value, C, is defined as $\bar{X}$ + 3 SD rounded to the nearest integer.

Table 19

Estimated Probability of the 03 Event, Mean, Standard Deviation, Critical Value, and Observed Frequencies in the 03 Cell for Data Set I, Three-Item Case, Over Two Replications (N=2319)

Objective Code	Replication 1					Replication 2				
	p_{03}[a]	$\bar{X}$	SD	C[b]	f_{03}	p_{03}	$\bar{X}$	SD	C	f_{03}
25-41	*.007*	0	4.01	12	17	*.004*	0	3.04	9	11
25-43	*.003*	0	2.63	8	7	*.004*	0	3.04	9	11
27-41	.00	0	0	0	0	.00	0	0	0	0
27-43	*.003*	0	2.63	8	8	*.003*	0	2.63	8	9
41-43	.41	951	23.68	1022	57	*.04*	0	9.43	28	89
22-24	.68	1577	22.46	1644	44	.70	1623	22.07	1689	46
22-25	*.01*	0	4.79	14	40	*.01*	0	4.79	14	36
22-27	*.01*	0	4.79	14	39	.66	1531	22.81	1599	41
22-41	.61	1415	23.49	1485	33	.00	0	0	0	0
22-43	*.01*	0	4.79	14	34	*.01*	0	4.79	14	40
24-25	*.009*	0	4.54	14	23	*.009*	0	4.54	14	22
24-27	*.01*	0	4.79	14	37	*.01*	0	4.79	14	36
24-41	.00	0	0	0	0	*.01*	0	4.79	14	40
24-43	*.007*	0	4.01	12	17	*.01*	0	4.79	14	26
25-27	*.003*	0	2.63	8	9	*.003*	0	2.63	8	9

[a] p_{03} is estimated from f_{03}/N when INCLU-value equalled 0. Estimated values are in italics.

[b] Critical value, C, is defined as $\bar{X}$ + 3 SD rounded to the nearest integer.

Table 20

Estimated Probability of the 02 Event, Mean, Standard Deviation, Critical Value and Observed Freqeuncies in the 02 Cell for Data Set IV, Two-Item Case, Over Three Replications (N=1118)

Objective Code	Replication 1					Replication 2					Replication 3				
	p_{02}[a]	$\bar{X}$	SD	C[b]	f_{02}	p_{02}	$\bar{X}$	SD	C	f_{02}	p_{02}	$\bar{X}$	SD	C	f_{02}
45-48	.02	22.36	4.68	36	19	*.01*	0	3.33	10	14	.01	11.18	3.33	21	13
45-55	.01	11.18	3.33	21	11	*.001*	0	1.06	3	2	.01	11.18	3.33	21	9
45-56	*.007*	0	2.78	8	8	.01	11.18	3.33	21	15	*.008*	0	2.98	9	9
48-55	.02	22.36	4.68	36	4	*.002*	0	1.49	4	3	.01	11.18	3.33	21	8
48-56	.04	44.72	6.55	64	32	.02	22.36	4.68	36	44	.04	44.72	6.55	64	57
55-56	.07	78.26	8.53	104	83	.05	55.90	7.29	78	95	.10	111.8	10.03	142	96

[a] p_{02} is estimated from f_{02}/N when INCLU-value equalled 0. Estimated values are in italics.

[b] Critical value, C, is defined as $\bar{X}$ + 3 SD rounded to the nearest integer.

Table 21

Estimated Probability of the 03 Event, Mean, Standard Deviation, Critical Value and Observed Frequencies in the 03 Cell for Data Set IV, Three-Item Case, Over Two Replications (N=1118)

Objective Code	Replication 1					Replication 2				
	p_{03}[a]	$\bar{X}$	SD	C[b]	f_{03}	p_{03}	$\bar{X}$	SD	C	f_{03}
45-48	*.009*	0	3.16	9	11	*.008*	0	2.98	9	9
45-55	*.007*	0	2.79	8	8	*.003*	0	1.83	5	4
45-56	*.008*	0	2.98	9	9	*.009*	0	3.16	9	11
48-55	*.002*	0	1.49	4	3	*.002*	0	1.49	4	3
48-56	*.02*	0	4.68	14	22	*.005*	0	2.36	7	6
55-56	.02	22.36	4.68	36	49	.01	11.18	3.33	21	54

[a] p_{03} is estimated from f_{03}/N when INCLU-value equalled 0. Estimated value are in italics.

[b] Critical value, C, is defined as $\bar{X}$ + 3 SD rounded to the nearest integer.

Table 22

Estimated Probability of the 02 Event, Mean, Standard Deviation, Critical Value and Observed Frequencies in the 02 Cell for Data Set V, Two-Item Case, Over Three Replications (N=1688).

Objective Code	Replication 1					Replication 2					Replication 3				
	p_{02}[a]	$\bar{X}$	SD	C[b]	f_{02}	p_{02}	$\bar{X}$	SD	C	f_{02}	p_{02}	$\bar{X}$	SD	C	f_{02}
55-56	.02	34	5.75	51	40	.01	17	4.09	29	20	.04	68	8.05	92	53
55-63	.13	219	13.82	260	110	.09	152	11.76	187	97	.03	51	7.01	72	29
55-64	.01	17	4.09	29	6	.05	84	8.95	111	53	.04	68	8.05	92	0
55-65	.09	152	11.76	187	84	.08	135	11.15	168	19	.03	51	7.01	72	26
56-63	.07	118	10.48	150	127	.02	34	5.75	51	69	.05	84	8.95	111	51
56-64	.04	68	8.05	92	0	.03	51	7.01	72	0	.01	17	4.09	29	0
56-65	.01	17	4.09	29	7	.05	84	8.95	111	17	.03	51	7.01	72	19
63-64	.01	17	4.09	29	0	.01	17	4.09	29	15	.04	68	8.05	92	9
63-65	.01	17	4.09	29	11	.03	51	7.01	72	3	.01	17	4.09	29	3
64-65	.03	51	7.01	72	94	.05	84	8.95	111	63	.18	304	15.78	351	139

[a] p_{02} is estimated from f_{02}/N when INCLU-value equaled 0.

[b] Critical value, C, is defined as $\bar{X}$ + 3 SD rounded to the nearest integer.

Table 23

Estimated Probability of the 03 Event, Mean, Standard Deviation, Critical Value, and Observed Frequencies in the 03 Cell for Data Set V, Three-Item Case, Over Two Replications (N=1688)

Objective Code	Replication 1					Replication 2				
	p_{03}[a]	$\bar{X}$	SD	C[b]	f_{03}	p_{03}	$\bar{X}$	SD	C	f_{03}
55-56	*.006*	10.13	3.17	20	10	.01	16.88	4.09	29	29
55-63	.01	16.88	4.09	29	40	.02	33.76	5.75	51	54
55-64	.01	16.88	4.09	29	0	.01	16.88	4.09	29	0
55-65	.02	33.76	5.75	51	12	*.002*	3.38	1.84	9	3
56-63	*.02*	33.76	5.75	51	31	*.04*	67.50	8.05	92	61
56-64	*.00*	0	0	0	0	.02	33.76	5.75	51	0
56-65	.01	16.88	4.09	29	5	*.008*	13.50	3.66	24	14
63-64	*.003*	5.06	2.25	12	5	*.004*	6.75	2.59	15	6
63-65	*.002*	3.38	1.84	9	3	*.002*	3.38	1.84	9	4
64-65	.25	422.00	17.79	475	2	.13	219.44	13.82	261	7

[a] p_{03} is estimated from f_{03}/N when INCLU-value equalled 0. Estimated values are in italics.

[b] Critical value, C, is defined as $\bar{X}$ + 3 SD rounded to the nearest integer.

of the "02 event," i.e., the probability that a member of the sample will be classified in the I-0, II-2 cell; the mean and standard deviation for the given distribution; a critical value "C," if when exceeded by the observed frequency in the 02 cell, will cause H_0 to be rejected; and the observed frequency, f_{02}, i.e., the actual number of cases classified in the I-0, II-2 cell.

In examining the f_{02} values for each objective-pair and suggesting tentative hierarchical relationships the following decision rules applied:

1. For any one objective-pair, H_0 was rejected if $f_{02} > C$; H_0 was accepted if $f_{02} \leq C$.
2. Over five replications, if H_0 was rejected 80% or more then no relationship in the objective-pair was acknowledged.
3. Objective-pairs were rank-ordered according to their percent of rejection, ranging from 0% to 60%.
4. Lower rejection-rate pairs were established in the hierarchy first followed by higher ones until all were accounted for.

The resulting hierarchies are displayed in Figures 8 through 10. In Data Set I, two objectives (24 and 25) were rejected from hierarchical consideration by decision rule one on two occasions. This resulted in two possible structures as the reader will note in Figure 8. In Data Set IV again

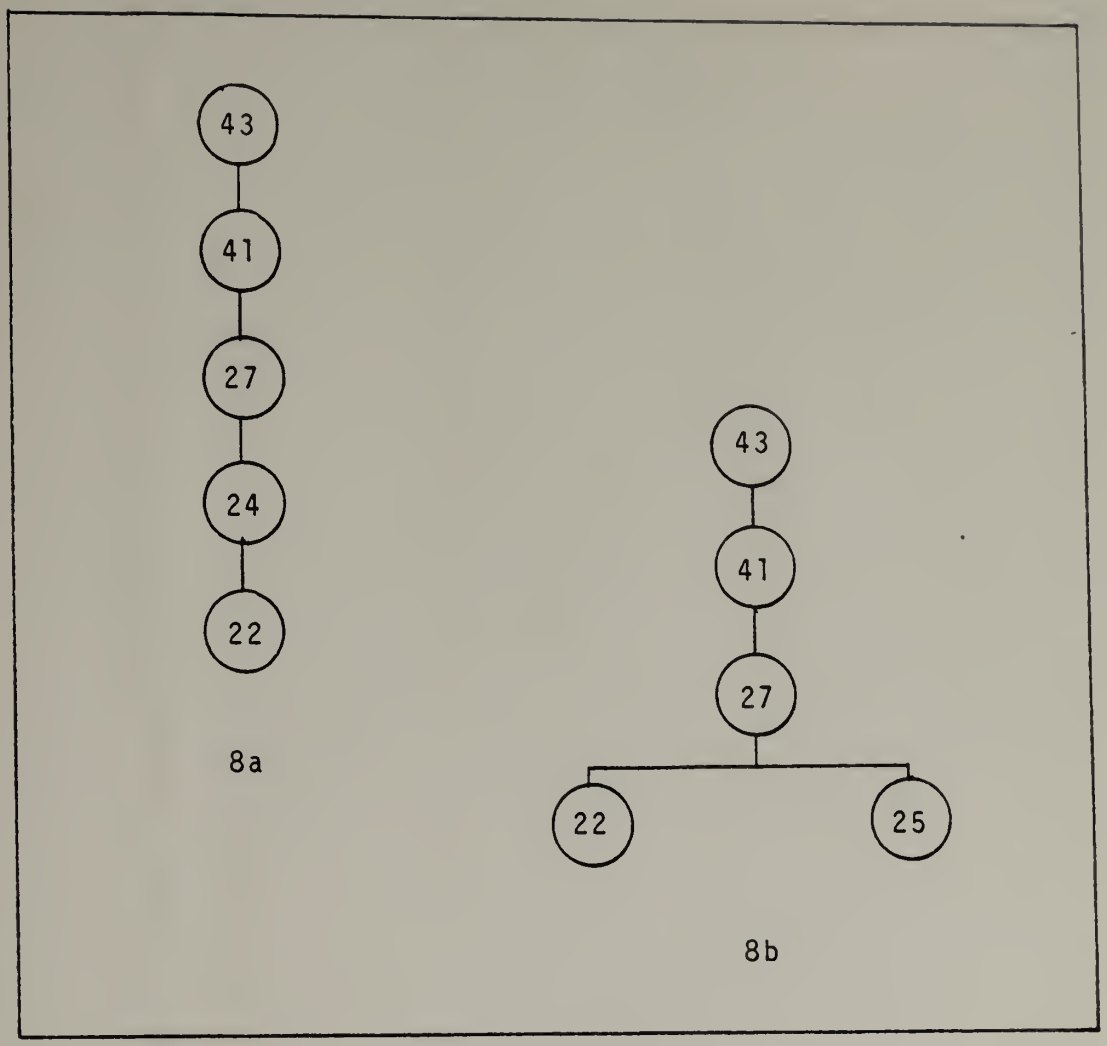

Figure 8. Proposed hierarchical structure for Data Set I based on the observed frequencies for each objective-pair found in Tables 18 and 19.

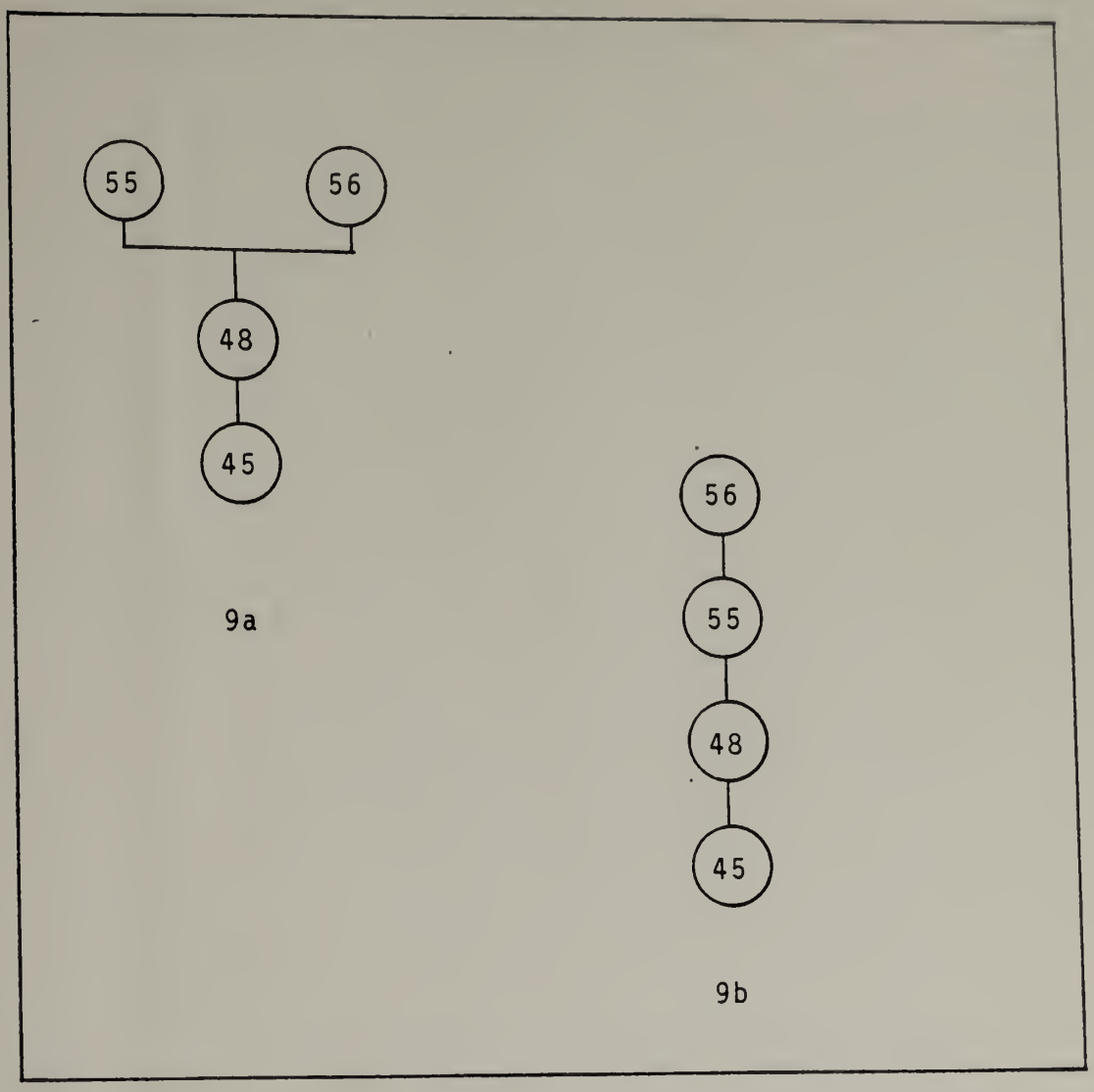

Figure 9. Proposed hierarchical structures for Data Set IV based on the observed frequencies for each objective-pair found in Tables 20 and 21.

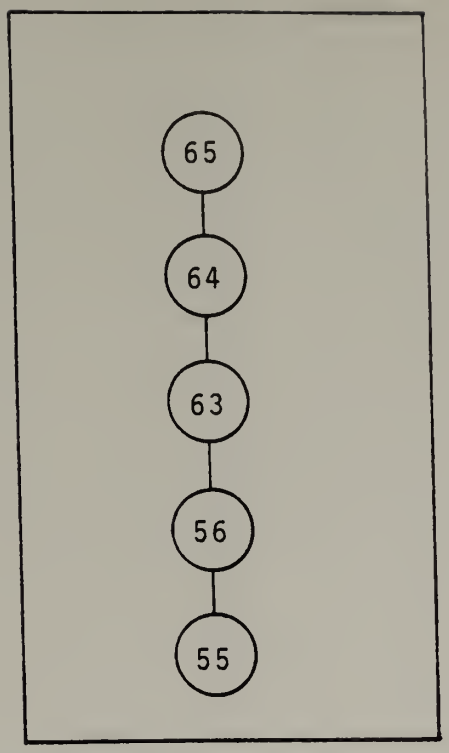

Figure 10. Proposed hierarchical structure for Data Set V based on the observed frequencies for each objective-pair found in Tables 22 and 23.

two objectives (55 and 56) while not rejected by decision rule one retained their position within the hierarchy with only weak supporting evidence. Therefore, it was deemed advisable to suggest two alternate structures displayed in Figure 9. Data Set V seemed to present clear evidence in favor of the structure provided in Figure 10.

4.2 Comparisons Across Methodologies

In the first section of this chapter we have presented the hierarchies resulting from three distinct methodological approaches for specifying and validating such hierarchies. The reader will note that the data have been presented in coded fashion with little or no regard for the content of each objective or objective-pair. At this point the author will compare and contrast all possible hierarchies for each of the data sets analyzed, but only after translating each one into its verbal analogue.

4.2.1 Data Set I

Data Set I is the largest of the three examined in this study, and thus, presented its own problems simply because of the total number of objectives in the sequence. The reader will note in Figure 11 that the hierarchies displayed vary in the total number of objectives included.

Those derived from "expert judgment" contain all six phonics objectives, with three variations on a theme. Those derived through the Dayton and Macready procedures omit ending consonant digraphs and/or beginning consonant digraphs. Those resulting from the application of the White and Clark model omit ending consonant sounds in one instance and auditory discrimination (rhyming) in the second case. There are some areas of agreement among the three. It appears that digraphs (beginning or ending or both) are at the top of the hierarchy, while consonant sounds (beginning or ending or both) are at the bottom. What comes in between is debatable, with vowel sounds and auditory discrimination (rhyming) as the potential candidates.

It seems also clear from the data that beginning sounds (either digraphs or consonants) precede ending sounds. The position of vowel sounds in the hierarchy is likewise worth noting. The position of this particular objective was clearly undecided by the experts and, if placed anywhere in the hierarchy, it was generally thought to be a branch of the primary hierarchy. This is quite contrary to the evidence provided by the two empirical models. The position of the auditory discrimination objective is somewhat surprising. One might assume that, as a listening skill, it falls into the pre-reading or readiness category, and as such, would have the lowest

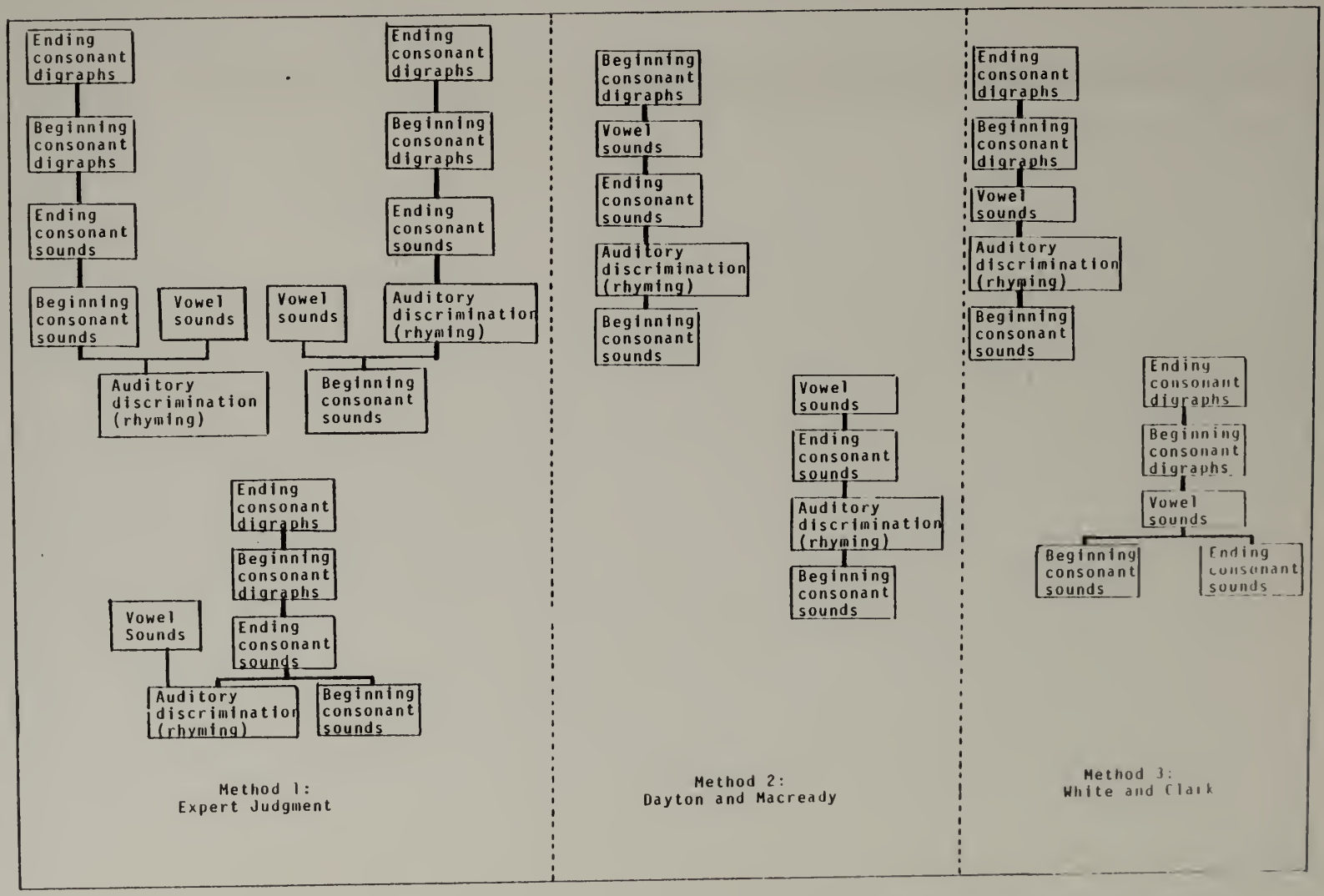

Figure 11. Proposed hierarchies for Data Set I resulting from three methodological approaches.

position in any hierarchy of formal reading skills. However, this does not appear to be the case. In both the judgment of the experts, as well as by the empirical evidence, this skill may be, and indeed is, preceded by beginning consonant sounds.

4.2.2 Data Set IV

Data Set IV has four objectives in total representing the structural analysis domain. Figure 12 presents the proposed hierarchical structures. The reader will notice that the expert judgment hierarchy omits the syllabication objective. Since the Dayton and Macready procedure failed to identify any relationships in this data set that were significant, this comparison will confine itself to the White and Clark model only. Clearly there is little evidence to support agreement between the a priori hierarchy and either of those resulting from the White and Clark procedures. Only the relationship between root word derivations and root words + affix seems clear, with the former lower in the hierarchy than the latter in both models.

Generally the failure of the Dayton and Macready procedure to identify any structures, coupled with the lack of agreement between White and Clark and expert judgment, would lead the author to suggest that no true

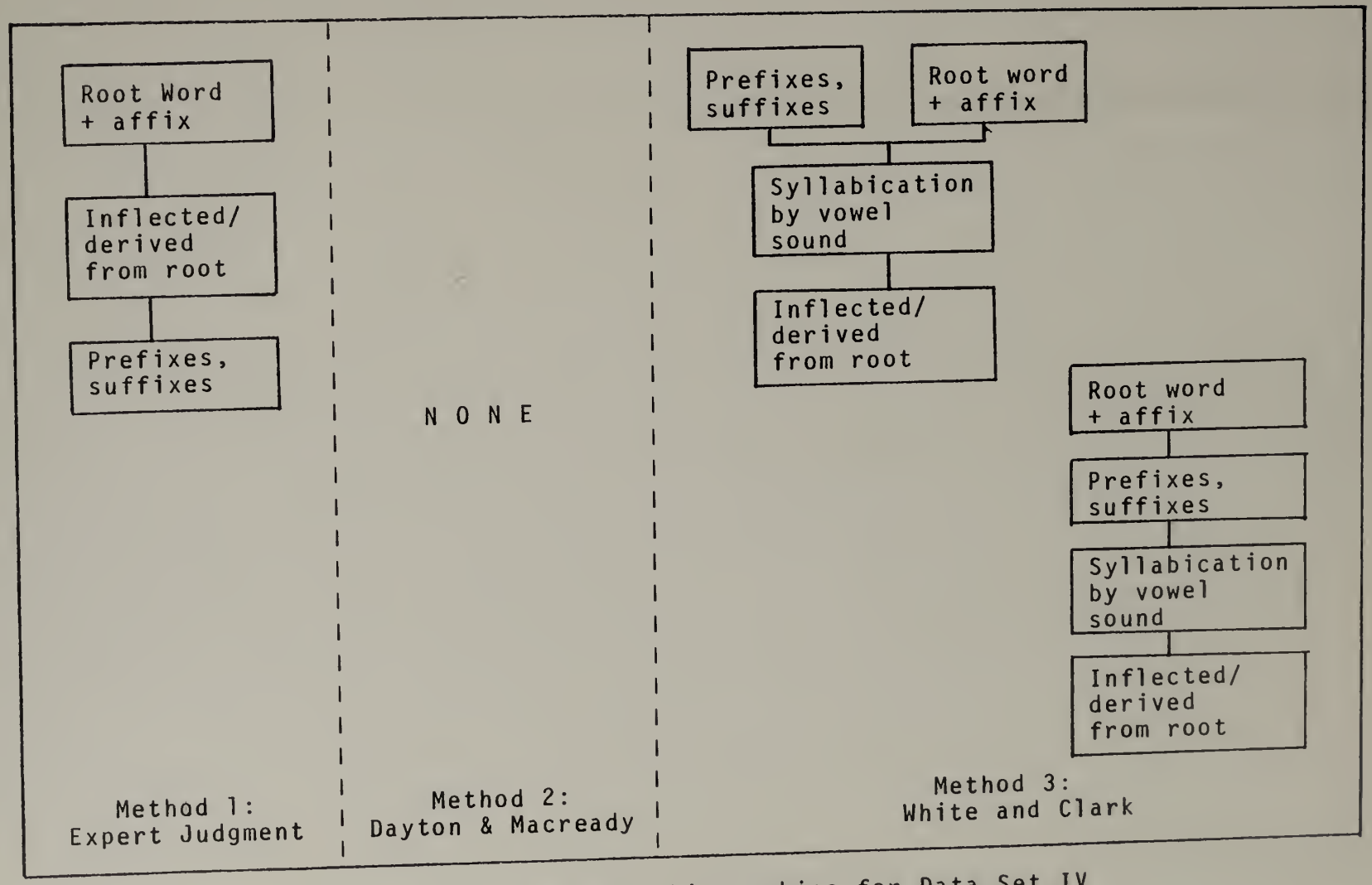

Figure 12. Proposed hierarchies for Data Set IV resulting from three methodological approaches.

hierarchy exists among this sequence of structural analysis skills.

4.2.3 Data Set V

This data set consisted of five objectives from the structural analysis domain. Figure 13 displays the resulting hierarchies suggested by the three procedures. The expert judgment hierarchy omits one objective from the sequence, viz., syllabication of nonsense words. Two are omitted from the Dayton and Macready sequence, viz., suffixes and syntax (verbs) and prefixes/suffixes. Only the White and Clark sequence maintains all five objectives. Notwithstanding the omissions, the degree of agreement among the three hierarchies is remarkably high. Two of the three agree that prefixes/suffixes is the lowest level skill of the five under consideration. All three agree that root word + affix is also a low level skill. Syllabication of nonsense words assumes the mid-position in two of the three hierarchies. Suffixes and syntax (nouns, verbs or both) top the list for all three models.

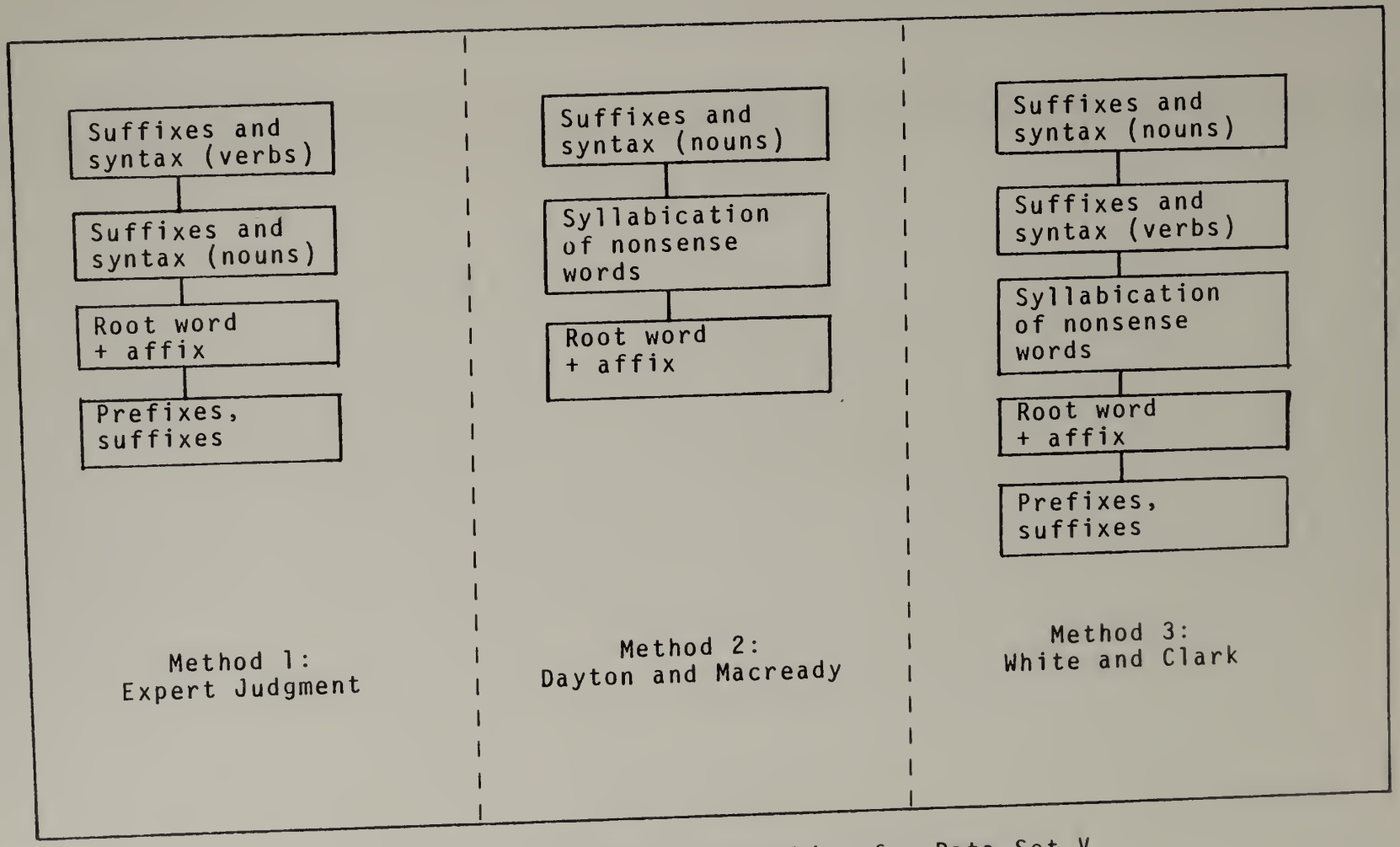

Figure 13. Proposed hierarchies for Data Set V resulting from three methodological approaches.

CHAPTER V

DISCUSSION AND CONCLUSIONS

5.1 Discussion of Results

The discussion of the results which follows will adhere to the same chronology as that used in the previous chapter. That is, each methodology will be discussed separately and then, final conclusions about the study as a whole will be reported.

5.1.1 The Expert Judgment Model

First, let us examine the hierarchy specification task performed by content experts. The basic question is, do experts in the field of reading admit of a hierarchy existing among a selected number of low-level phonics and structural analysis skills. Generally, the data seem to indicate that the experts concur on only one aspect of hierarchical relationships, viz., sequence. That is, the data indicate a willingness to say Skill I precedes Skill II. However, they stop short of saying that Skill I *must* be mastered before Skill II *can* be mastered; or that the mastery of Skill II will be facilitated by the prior mastery of Skill I; or the mastery of both skills is mutually dependent, one on the other.

On an informal basis some of the respondents were interviewed after completion of the hierarchy specification task. Several acknowledged that they found the objectives "fuzzy," i.e., not articulated in behavioral terms, and thus, were forced to make arbitrary decisions about the pairwise relationships. Others openly acknowledged their personal biases with regard to reading hierarchies. Exponents of this particular philosophy claim that the process of learning to read is unique to each learner, and thereby, defies the sequencing of reading skills in a particular way on a universal basis. This is what the author refers to as the "different strokes" school of thought. Finally, there were those who questioned the idea of "mastery" saying that the acquisition of reading skills is a cyclic process that includes skill instruction, reinforcement, application, further skill instruction, more reinforcement, new applications, and so forth. Exponents of this school of thought claim that "mastery" is a penultimate event occurring just prior to the state of being a "reader," and not a series of discrete events along the instructional path of "learning to read."

In the final analysis, the hierarchies resulting from the judgment of experts were at best weak, and at worst forced. Although the author has suggested some hierarchical structures, they should be viewed cautiously.

5.1.2 The Dayton and Macready Model

The hypothesis of interest was

$$H_0: p_{j_{obs}} = p_{j_{pred}}$$

$$H_1: p_{j_{obs}} \neq p_{j_{pred}}$$

Clearly, the null hypothesis must be rejected based on the chi-square results given in Table 17. None of the data sets lends itself to a hierarchical interpretation based on a linear sequence of true score patterns. Examination of the values of α and β shows that only for Data Set I are these parameters within expected bounds. The unusually high value of the "forgetting" parameter in Data Sets IV and V can possibly be interpreted in several ways. First, and the most obvious, is that a linear sequence just does not provide a reasonable fit for the data. A second possible interpretation is that examinees were not appropriately assigned to consecutive test levels. While this interpretation is not so obvious, it is, in the opinion of the author, one of the most likely reasons for the apparent lack of fit of the data. Further substantiating evidence is provided for this interpretation by the relatively high values of the "guessing" parameter as well. Also, it should be noted that as the criterion score for mastery was lowered from n-1 to n-2, the values for both parameters began to approach more reasonable limits. And

when the criterion score is a composite of two or three items, as in the White and Clark model, the degree of fit is improved markedly. One might conclude, therefore, that the fewer the items the better the fit, which in this case, is quite true. However, this is contrary to what we already know about the assessment of mastery and the required number of items necessary to minimize the number of misclassifications.

5.1.3 The White and Clark Model

The hypothesis of interest in this model was

$$H_0: f_{on} \leq C$$

$$H_1: f_{on} > C$$

where f_{on} has been previously defined as the observed frequency in the I-0, II-2 cell in the two-item case and I-0, II-3 cell in the three-item case; and C is a selected threshold value, which, if exceeded, would exclude Skill I and/or II from the hierarchy under consideration. In this particular study the value of C was further defined as the nearest integer point to the third standard deviation above the mean of the sampling distribution. It should be noted that the normal approximation to the binomial distribution was used in order to simplify the estimation of the distributional statistics. Because of the multiple

replications it was necessary to establish the five decision rules defined in the preceding chapter. On that basis it was possible to arrive at the hierarchical structures displayed in Figures 11, 12 and 13. This model is really a special case of the general probabilistic model, as pointed out by Dayton and Macready (1967a), which utilizes marginal frequencies instead of a maximum likelihood estimation procedure.

The application of the White and Clark model appears to be less sensitive to some of the overall distributional problems found in the Dayton and Macready approach. Consequently, the hierarchical structures emerged without undue strain on the data. Now, of course, at least two assumptions were made by the researcher in addition to those indigenous to the model itself: (1) that each replication was an equally reliable estimate of the probability of the "02 event"; and (2) that all items were equally reliable measures of mastery of the consequent skill. Both assumptions seem reasonable in view of the fact that test items were generated on the basis of domain-specifications, and that their order of presentation to the examinee was random and not according to item-difficulty indices. In fact, the resulting structures match fairly well with the a priori hierarchies generated by the content experts.

5.1.4 Other Results

Three other informal hypotheses were proposed in Chapter III.

That the Dayton and Macready (1976a, 1976b) procedure is the preferred one when the sample size is large is not clear from the evidence presented here. The author has suggested that a large N has led to the rejection of H_0. However, one would need to examine several smaller samples to see if indeed this is true.

The author hypothesized that multi-item sets would be superior to single-item estimations in establishing response patterns. Because of the limitations of the testing program design just the opposite appeared to be the case. This may have been an artifact, however.

Finally, the superiority of the maximum likelihood estimation procedures over those utilizing marginal frequencies was postulated. Although theoretically this is usually true, in this particular study no further evidence to support such a posture was immediately evident.

5.2 Limitations of the Study

There were several limitations in this study which to some extent were outside the control of the researcher. Except in the most unusual of circumstances, this is nearly always the case in field-based research.

First of all, the curriculum objectives which formed the basis of the seqeunce of skills under consideration were not articulated in behavioral terms. For example:

Curriculum Objective as Stated:	*The learner knows the consonant blends or digraphs he/she hears at the beginning of two dictated words.*
Curriculum Objective Stated in Behavioral Terms:	*The student selects the illustration representing a word that begins with the same consonant digraph as the oral stimulus word.*

If each of the objectives in the sequence had been articulated more precisely the confusion in completing the hierarchy specification task by the experts may have been ameliorated to a large extent.

Secondly, the item statistics on the *Inventory* indicated that some items were not reliable measures of the objective. Consequently, a revision was in order. However, as noted earlier, at the time of testing the district was still using an unrevised pilot edition of the test. Unreliability of the instrument would undoubtedly lead to erroneous conclusions about the hierarchical structure (or lack of it).

Thirdly, the design of the testing program itself is open to several limitations which impacted upon this study. The assignment of examinees to test levels was based on pre-test scores generated twelve months prior to this data-collection testing period. During the intervening

period the assumption was that instruction would be directed to teaching those skills to be assessed for mastery at post-test time. There are no guarantees that such direct instruction indeed did occur. Additionally, the assignment of test levels precluded measuring many or all of the objectives in a sequence since no student took more than two contiguous levels, and some took only one level. This severely limited the total number of objectives assessed for any one examinee.

Finally, the nature of the *Inventory* is summative. That is, it measures only selected skills from a more comprehensive sequence of skills. Those which were included were arbitrarily judged to be appropriate for inclusion because they were "the most difficult" skills by some internal standard known only to district staff.

5.3 Suggestions for Further Research

It is clear from a consideration of the results of this study that White's procedure (1974b) indeed outlines the optimal steps for conducting hierarchy research under properly controlled conditions. Given that those conditions exist it should be possible to compare several alternate hierarchies fitted to the same data. For example, in this study a strictly linear hierarchy was examined. However, it is possible under the Dayton and Macready (1976a, 1976b) model to consider branching

hierarchies as well. The focus of any further research in this area should be directed at a rigidly controlled design that precludes the invalidating features of a field-based approach. Within the constraints of such a design it would be quite possible to examine aspects of item-design and its relation to hierarchy validation.

One might also wish to examine the long-term effects of hierarchical learning. To date, only a few studies have examined the issue of retention of hierarchically learned skills over a period of time (White, 1976d and White & Gagné, 1978). Both of these studies postulate that if learning is hierarchical so also is retention. Whether or not this hypothesis can be applied in the area of reading needs to be explored.

This study has shown that quite possibly learning hierarchies do exist in the area of reading. Future research should focus on limited content domains to affirm or disaffirm this initial evidence.

REFERENCES

Airasian, P. W. A method for validating sequential instructional hierarchies. *Educational Technology*, 1971, *11*, 54-56.

Airasian, P. W., & Bart, W. M. Ordering theory: A new and useful measurement model. *Educational Technology*, 1973, *13*, 56-60.

Airasian, P. W., & Bart, W. M. Validating a priori instructional hierarchies. *Journal of Educational Measurement*, 1975, *12*, 163-173.

Baker, F. B., & Hubert, L. J. Inference procedures for ordering theory. *Journal of Educational Statistics*, 1977, *2*, 217-233.

Bart, W. M., & Airasian, P. W. Determination of the ordering among seven Piagetian tasks by an ordering-theoretic method. *Journal of Educational Psychology*, 1974, *66*, 277-284.

Bart, W. M., & Krus, D. J. An ordering-theoretic method to determine hierarchies among items. *Educational and Psychological Measurement*, 1973, *33*, 291-300.

Campbell, D. T., & Stanley, J. C. *Experimental and quasi-experimental designs for research*. Chicago: Rand-McNally & Co., 1966.

Capie, W., & Jones, H. L. An assessment of hierarchy validation techniques. *Journal of Research in Science Teaching*, 1971, *8*, 137-147.

Cotton, J. W., Gallagher, J. P., & Marshall, S. P. The identification and decomposition of hierarchical tasks. *American Educational Research Journal*, 1977, *14*, 189-212.

Cronbach, L. J. Test validation. In R. L. Thorndike (Ed.), *Educational measurement*. (2nd ed.) Washington, D.C.: American Council on Education, 1971.

Dayton, C. M., & Macready, G. B. A probabilistic model for validation of behavioral hierarchies. *Psychometrika*, 1976, *41*, 189-204. (a)

Dayton, C. M., & Macready, G. B. *Computer programs for probabilistic models.* Unpublished manuscript, 1976. (Available from the authors, University of Maryland, College Park, Maryland.) (b)

Dzuibon, C. C., & Vickery, K. V. Criterion-referenced measurement: Some recent developments. *Educational Leadership*, 1973, *30*, 483-486.

Gagné, R. M. The acquisition of knowledge. *Psychological Review*, 1962, *69*, 355-365.

Gagné, R. M. Learning hierarchies. *Educational Psychologist*, 1968, *6*, 1-9.

Gagné, R. M. *The conditions of learning.* (2nd ed.) New York: Holt, Rinehart, and Winston, 1970. (a)

Gagné, R. M. Some new views of learning and instruction. *Phi Delta Kappan*, 1970, *51*, 468-472. (b)

Gagné, R. M., Mayor, J. R., Garstens, H. L., & Paradise, N. E. Factors in acquiring knowledge of a mathematical task. *Psychological Monographs*, 1962, *76*, (7, Whole No. 526).

Gagné, R. M., & Paradise, N. E. Abilities and learning sets in knowledge acquisition. *Psychological Monographs*, 1961, *75*, (14, Whole No. 518).

Glaser, R., & Nitko, A. J. Measurement in learning and instruction. In R. L. Thorndike (Ed.), *Educational measurement.* (2nd ed.) Washington, D.C.: American Council on Education, 1971.

Hambleton, R. K. Testing and decision-making procedures for selected individual instructional programs. *Review of Educational Research*, 1974, *44*, 371-400.

Hambleton, R. K. *The reading skills inventory: A criterion-referenced assessment.* Unpublished pilot edition, 1975.

Hambleton, R. K., & Eignor, D. R. *Adaptive testing applied to hierarchically structured objective-based programs.* In D. J. Weiss (Ed.), *Proceedings of the 1977 Computerized Adaptive Testing Conference.* Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1978.

Hambleton, R. K., Hutten, L. R., & Swaminathan, H. A comparison of several methods for assessing student mastery in objectives-based instructional programs. *Journal of Experimental Education*, 1976, *45*, 57-64.

Hambleton, R. K., & Novick, M. R. Toward an integration of theory and method for criterion-referenced tests. *Journal of Educational Measurement*, 1973, *10*, 159-170.

Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. Criterion-referenced testing and measurement: A review of technical issues and developments. *Review of Educational Research*, 1978, *47*, 1-47.

Harris, C. W., Alkin, M. C., & Popham, W. J. (Eds.) *Problems in criterion-referenced measurement*. CSE Monograph Series in Evaluation, No. 3. Los Angeles: Center for the Study of Evaluation, University of California, 1974.

Hively, W., Patterson, H. L., & Page, S. A. A "universe-defined" system of arithmetic achievement tests. *Journal of Educational Measurement*, 1968, *5*, 275-290.

Hsu, T. *Empirical data on criterion-referenced tests*. Pittsburgh, PA.: University of Pittsburgh, Learning Research and Development Center, 1971. (ERIC Document Reproduction Service No. ED 050 139.)

Huey, E. B. *The psychology and pedagogy of reading*. Cambridge, MA: MIT Press, 1968. (Originally published, 1908.)

Larrivee, B. Behavioral objective checklist for reading and related skills. Unpublished Manuscript, 1977.

Lingoes, J. C. Multiple scalogram analysis: A set theoretic model for analyzing dichotomous items. *Educational and Psychological Measurement*, 1963, *23*, 3.

Linke, R. D. Replicative studies in hierarchical learning of graphical interpretation skills. *British Journal of Educational Psychology*, 1975, *45*, 39-46.

Macready, G. B. The structures of domain hierarchies found within a domain referenced testing system. *Educational and Psychological Measurement*, 1975, *35*, 583-598.

Macready, G. B., & Dayton, C. M. The use of probabilistic models in the assessment of mastery. *Journal of Educational Statistics*, 1977, *2*, 99-120.

Macready, G. B., & Merwin, J. C. Homogeneity within item forms in domain referenced testing. *Educational and Psychological Measurement*, 1973, *33*, 351-360.

Millman, J. Passing scores and test lengths for domain-referenced tests. *Review of Educational Research*, 1973, *43*, 205-216.

Niesser, U. *Cognitive psychology*. New York: Appleton-Century-Crofts, 1967.

Rao, C. R. *Linear statistical inference and its applications*. New York: John Wiley, 1965.

Resnick, L. B., & Wang, M. C. Approaches to the validation of learning hierarchies. *Proceedings of the Eighteenth Annual Regional Conference on Testing Problems*. Princeton, New Jersey: Educational Testing Service, 1969.

Resnick, L. B., Wang, M. C., & Kaplan, J. Task analysis in curriculum design: A hierarchically sequenced introductory mathematics curriculum. *Journal of Applied Behavior Analysis*, 1973, *6*, 679-710.

Swaminathan, H., Hambleton, R. K., & Algina, J. Reliability of criterion-referenced tests: A decision-theoretic formulation. *Journal of Educational Measurement*, 1974, *11*, 263-268.

Swaminathan, H., Hambleton, R. K., & Algina, J. A Bayesian decision-theoretic procedure for use with criterion-referenced tests. *Journal of Educational Measurement*, 1975, *12*, 87-98.

Torgerson, W. S. *Theory and methods of scaling*. New York: John Wiley, 1958.

Walbesser, A. H., & Eisenberg, T. A. *A review of research on behavioral objectives and learning hierarchies*. Columbus, Ohio: ERIC Information Analysis Center for Science, Mathematics and Environmental Education, 1972. (ERIC Document Reproduction Service No. ED 059 900.)

White, R. T. Research into learning hierarchies. *Review of Educational Research*, 1973, *43*, 361-375.

White, R. T. The validation of a learning hierarchy. *American Educational Research Journal*, 1974, *11*, 121-136. (a)

White, R. T. A model for the validation of learning hierarchies. *Journal of Research in Science Teaching*, 1974, *11*, 1-3. (b)

White, R. T. Indexes used in testing the validity of learning hierarchies. *Journal of Research in Science Teaching*, 1974, *11*, 61-66. (c)

White, R. T. Effects of guidance, seqeunce and attribute-treatment interactions on learning, retention, and transfer of hierarchically ordered skills. *Instructional Science*, 1976, *5*, 133-152. (d)

White, R. T., & Clark, R. M. A test of inclusion which allows for errors of measurement. *Psychometrika*, 1973, *38*, 77-86.

White, R. T., & Gagné, R. M. Formative evaluation applied to a learning hierarchy. *Contemporary Educational Psychology*, 1978, *3*, 87-94.

Wolf, T. Reading reconsidered. *Harvard Educational Review*, 1977, *47*, 411-429.

APPENDIX A

Content Objectives of the *Reading Skills Inventory: A Criterion-Referenced Assessment*[1]

[1]The objectives were prepared by staff and teachers in the Department of Public Schools of a small, urban community in New England.

Appendix A

Content Objectives of Reading Skills Inventory by Category

Curriculum Code	Objective
	Visual-Perceptual-Motor Skills (VPM)
VPM-01	The learner knows how to arrange several objects in a specific order, according to size.
VPM-02	The learner knows how to classify several objects into two groups, according to fucntion.
VPM-08	The learner knows how to fill in the missing parts of letters and numbers that are incomplete, when accompanied by the complete form.
VPM-09	The learner knows how to arrange three letters in a specific order.
VPM-10	The learner knows how to connect matching letters with a line.
VPM-11	The learner knows that one word is different from three others in a group of four words.
	Listening-Oral Language Skills (LOL)
LOL-04	The learner knows words dealing with the home situation.
LOL-07	The learner knows words dealing with the neighborhood situation.
	Pre-Reading-Readiness Skills (PRR)
PRR-01	Given groups of letters, the learner can select the letter the teacher dictates.
PRR-02	The learner reproduces the letters of the alphabet as dictated by the teacher.

Curriculum Code	Objective
PRR-03	Given a dictated word, the learner marks the picture with the same initial sound.
PRR-05	From a list of choices, the learner marks the word that is the same as the first word.
	Phonics Skills (PH)
PH-01	The learner knows the consonant corresponding to the sounds he/she hears at the beginning of of two dictated words.
PH-02	The learner knows the sound of a given consonant and matches it to a picture beginning with the same sound.
PH-03	The learner knows how to make new words by substituting initial consonants in known words.
PH-05	The learner knows the consonant corresponding to the sound he/she hears at the end of two dictated words.
PH-06	The learner knows whether a given consonant sound is at the beginning, middle, or end of a dictated word.
PH-07	The learner knows the consonant blend or digraph he/she hears at the beginning of two dictated words.
PH-08	The learner knows how to substitute initial consonant blends and digraphs in words.
PH-09	The learner knows the consonant blend or digraph he/she hears at the end of two dictated words.
PH-10	The learner knows the vowels heard in a dictated word.
PH-11	The learner knows the long and short vowel sounds.

Curriculum Code	Objective
PH-12	The learner knows the long, short and "r" controlled vowels.
PH-14	The learner knows the vowels he/she hears in a dictated word of one or more syllables. (less difficult than PH-22)
PH-15	Given a list of the most common vowel principles, the learner can apply the principles properly.
PH-16	The learner knows how to apply the vowel principles properly to a nonsense word.
	Structural Analysis Skills (SA)
SA-01	The learner can identify the simple endings that denote tense, number, person, possession and comparison.
SA-02	The learner knows the root word in an inflected form or in a derived form.
SA-03	The learner knows how to divide a compound word into its component parts.
SA-04	The learner knows the two words indicated in a contraction.
SA-05	The learner knows how to identify prefixes and suffixes in a list of derivatives.
SA-06	Given a list of prefixes and suffixes, the learner knows the affix to be added to a given root word to make sense in a sentence.
SA-07	The learner knows the number of syllables heard in a word by counting the vowel sounds.
SA-08	The learner applies the vowel principles to syllables and can indicate whether the vowel sound in the first syllable is long, short or controlled.

Curriculum Code	Objective
SA-09	Given a list of two syllable nonsense words, the learner knows how to divide them into syllables according to the principles of syllabication.
SA-10	Given a list of two syllable nonsense words, the learner knows how to pronounce them.
SA-11	Given a choice of suffixes, the learner can select the one which changes the function of a word to become a word that shows action.
SA-12	Given a choice of suffixes, the learner can select the one which changes the meaning of a word to become a word which names.
	Dictionary Skills (DS)
DS-01	The learner knows alphabetical order.
DS-02	The learner knows how to alphabetize by the first letter.
DS-03	The learner knows how to alphabetize by the second and third letter.
DS-04	The learner knows how to use guide words in the dictionary.
DS-06	The learner applies the appropriate dictionary meaning to fit the context.

APPENDIX B

Hierarchy Specification Forms

For Selected Objectives From

The Reading Skills Inventory

DIRECTIONS

The purpose of this task is to determine the relationship, if any, among a selected set of word attack skills that are common to most reading programs. On the reverse side there is a list of skills divided into two categories, SET I and SET II. Each SET will be reviewed as a separate unit. Since there are nine (9) skills in each SET, there will be 36 pairs of skills to review in each SET. The list gives a complete statement of the skill as well as the abbreviated version which will appear as you examine each pair.

TASK

Your task is to examine each pair as they appear on the enclosed sheets and to respond to two questions:

Q. Is learning one skill in the pair necessary to learning the alternate skill listed? If so, which must be learned first?

Q. Will learning one skill in the pair facilitate learning the alternate skill listed? If so, which should be learned first?

Thank you for your cooperation in this project. Your time is responding is greatly appreciated.

You may return your response package in the envelope provided at your earliest convenience.

Sample Task Form

I-1-3

A. *Beginning consonant sounds*

B. *Ending consonant sounds*

( ) Learning one skill is NECESSARY to learn the other:
- ( ) A before B
- ( ) B before A

( ) Learning one skill will FACILITATE learning the other:
- ( ) A before B
- ( ) B before A
- ( ) Each facilitates the other

( ) The skills can effectively be learned in either order, A before B or B before A, since order of learning is irrelevant.

Sample Task Form

I-2-1

A. *Auditory discrimination: rhyming*

B. *Beginning consonant sounds*

( ) Learning one skill is NECESSARY to learn the other:
- ( ) A before B
- ( ) B before A

( ) Learning one skill will FACILITATE learning the other:
- ( ) A before B
- ( ) B before A
- ( ) Each facilitates the other

( ) The skills can effectively be learned either order, A before B or B before A, since order of learning is irrelevant.

SET I

Skill Statement	Abbreviated Version
1. The learner knows the consonant corresponding to the sounds he/she hears at the beginning of two dictated words.	1. *Beginning consonant sounds.*
2. The learner knows how to make new words by substituting initial consonants in known words.	2. *Auditory discrimination: rhyming.*
3. The learner knows the consonant corresponding to the sound he/she hears at the end of two dictated words.	3. *Ending consonant sounds.*
4. The learner knows the consonant blend or digraph he/she hears at the beginning of two dictated words.	4. *Beginning consonant digraphs.*
5. The learner knows the consonant blend or digraph he/she hears at the end of two dictated words.	5. *Ending consonant digraphs.*
6. The learner knows the vowels heard in a dictated word.	6. *Vowel sounds.*
7. The learner knows the long, short, and "r" controlled vowels.	7. *Long/short/r-controlled vowels.*
8. The learner knows alphabetical order.	8. *Alphabetical order.*
9. The learner knows how to apply the vowel principles properly to a nonsense word.	9. *Application vowel principles to nonsense words.*

SET II

Skill Statement	Abbreviated Version
1. The learner can identify the simple endings that denote tense, number, person, possession and comparison.	1. *Suffixes denoting syntax.*
2. The learner knows the root word in an inflected form or in a derived form.	2. *Inflected/derived from root word.*
3. The learner knows how to identify prefixes and suffixes in a list of derivatives.	3. *Prefixes, suffixes.*
4. Given a list of prefixes and suffixes, the learner knows the affix to be added to a given root word to make sense in a sentence.	4. *Root word + affix.*
5. The learner knows the number of syllables heard in a word by counting the vowel sounds.	5. *Syllabication by vowel sound.*
6. The learner knows how to alphabetize by the first letter.	6. *Alphabetize by first letter.*
7. Given a list of two syllable nonsense words, the learner knows how to divide them into syllables according to the principles of syllabication.	7. *Syllabication of nonsense words.*
8. Given a choice of suffixes, the learner can select the one which changes the function of a word to become a word that shows action.	8. *Suffixes and syntax (verbs).*
9. Given a choice of suffixes, the learner can select the one which changes the meaning of a word to become a word which names.	9. *Suffixes and syntax (nouns).*